Zubair *et al. Journal of Statistical Distributions and Applications* (2018) 5:1
DOI 10.1186/s40488-017-0082-3

Journal of Statistical Distributions and Applications

RESEARCH

Open Access

The power-Cauchy negative-binomial: properties and regression

Muhammad Zubair[1], Muhammad H. Tahir[2], Gauss M. Cordeiro[3*], Ayman Alzaatreh[4] and Edwin M. M. Ortega[5]

*Correspondence:
gausscordeiro@gmail.com
[3]Departamento de Estatística, Universidade Federal de Pernambuco, PE 50740-540, Recife, Brazil
Full list of author information is available at the end of the article

Abstract

We propose and study a new compounded model to extend the half-Cauchy and power-Cauchy distributions, which offers more flexibility in modeling lifetime data. The proposed model is analytically tractable and can be used effectively to analyze censored and uncensored data sets. Its density function can have various shapes such as reversed-J and right-skewed. It can accommodate different hazard shapes such as decreasing, upside-down bathtub and decreasing-increasing-decreasing. Some mathematical properties of the new distribution can be determined from a linear combination for its density function such as ordinary and incomplete moments. The performance of the maximum likelihood method to estimate the model parameters is investigated by a simulation study. Further, we introduce the new log-power-Cauchy negative-binomial regression model for censored data, which includes as sub-models some widely known regression models that can be applied to censored data. Four real life data sets, of which one is censored, have been analyzed and the new models provide adequate fits.

Keywords: Censoring, Compounding, G-class, Half-Cauchy distribution, Maximum likelihood estimation, Negative-binomial distribution

AMS Subject Classification: Primary 60E05, Secondary 62N05, 62F10

Introduction

Numerous extended classical distributions have been proposed for modelling data in several areas such as biological studies, environmental and medical sciences, engineering, economics, finance and actuarial science. However, in many applied areas like lifetime analysis, finance and insurance, there is a clear need for further extended distributions, that is, new distributions which are more flexible to model real data in these areas, since the data can present a high degree of skewness and kurtosis. There are many generalizations and extensions of distributions in literature using the randomly-stopped approach for either the minimum or maximum of K independent and identically distributed (iid) random variables (discrete or continuous). See, for example, Nekoukhou and Bidram (2017). Further, Rooks et al. (2010) introduced a two-parameter *power-Cauchy (PC)* distribution for analyzing upside-down bathtub (UBT) hazard function data. The cumulative distribution function (cdf) and probability density function (pdf) of the PC distribution with shape parameter α and scale parameter σ are, respectively, given by

$$G_{PC}(z;\alpha,\sigma) = 2\,\pi^{-1}\,\tan^{-1}(z/\sigma)^{\alpha}\,, \quad z>0 \quad \alpha,\sigma>0 \tag{1}$$

and

$$g_{PC}(z;\alpha,\sigma) = 2\,\pi^{-1}\,(\alpha/\sigma)\,(z/\sigma)^{\alpha-1}\left[1+(z/\sigma)^{2\alpha}\right]^{-1}. \tag{2}$$

Tahir et al. (2016) studied the *exponentiated power-Cauchy (EPC)* distribution. Let Z_a denote the EPC distribution with baseline parameters α and σ and power parameter $a>0$. The cdf and pdf of Z_a are given by

$$F_{EPC}(z) = \left[2\pi^{-1}\,\tan^{-1}\left(\frac{z}{\sigma}\right)^{\alpha}\right]^{a} \tag{3}$$

and

$$f_{EPC}(z) = 2\,a\,\pi^{-1}\left(\frac{\alpha}{\sigma}\right)\left(\frac{z}{\sigma}\right)^{\alpha-1}\left[1+\left(\frac{z}{\sigma}\right)^{2\alpha}\right]^{-1}\left[2\pi^{-1}\,\tan^{-1}\left(\frac{z}{\sigma}\right)^{\alpha}\right]^{a-1}, \tag{4}$$

respectively.

In this paper, we define a new four-parameter generalization of the PC distribution named the *power-Cauchy negative-binomial* (PCNB) model. The new distribution is flexible to model complex positive real data sets, i.e., it can have decreasing, UBT shaped and decreasing-increasing-decreasing hazard rate functions (hrfs). It thus provides a good alternative to several well-known life distributions.

The paper is unfolded as follows. In “The proposed model” section, we define the PCNB distribution. In “Properties of the new model” section, we obtain some of its mathematical properties including quantile function (qf), tail behaviors, a useful linear representation for its density function and some types of moments. In “Estimation” section, the model parameters are estimated by maximum likelihood and a simulation study is performed. In “Regression model” section, we present a regression model based on the PCNB distribution with censored data. In “Applications” section, the usefulness of the new distribution is illustrated by means of four real data sets where we show empirically that it outperforms some well-known lifetime distributions. Finally, “Concluding remarks” section offers some concluding remarks.

The proposed model

General Insurance companies typically face two major problems when they want to use past or present claim amounts in forecasting future claim severity. First, they have to find an appropriate statistical distribution for their large volumes of claim amounts. Then, test how well this statistical distribution fits their claim data. Most data in general insurance problems is skewed to the right and therefore most distributions that exhibit this characteristic can be used to model the claim severity. Insurance data contains relatively large claim amounts, which may be infrequent. Hence, there is a clear need to use statistical distributions with relatively heavy tails and highly skewed like the PC distribution.

Large claims play a special role because of their importance financially. It is also hard to assess their distribution. They do not occur very often, and historical experience is therefore limited. Insurance companies may even cover claims larger than anything that has been seen before. How should such situations be tackled? The simplest would be to fit a parametric family and try to extrapolate beyond past experience. That may not be a very good idea. A generalization of the PC distribution may fit well in the central regions without being reliable at all at the extreme right tail, and such a procedure may easily underestimate big claims severely.

Let $T_1, \ldots, T_K$ denote the failure times of K (a latent random variable) claims where K is assumed independent of the T_i's in a set-up with at least one claim. Then, we define $Z = \max\{T_1, \ldots, T_K\}$. Consider that the T_i's are iid random variables with common cdf $G(z)$ and that K follows the negative-binomial (NB) probability mass function ($n = 1, 2, \ldots$ and p are fixed but unknown parameters)

$$\mathbf{P}(K = k) = \binom{k-1}{n-1} p^n \, (1-p)^{k-n}, \; k = n, n+1, \ldots, \quad p \in (0,1).$$

Under this set-up, the conditional pdf of Z given K is

$$f(z \mid K = k) = k\, g(z)\, G(z)^{k-1}.$$

Then, the marginal pdf of Z follows as

$$\begin{aligned} f(z) &= \sum_{k=n}^{\infty} k\, g(z)\, G(z)^{k-1} \binom{k-1}{n-1} p^n \, (1-p)^{k-n} \\ &= \frac{p^n g(z)}{(1-p)^n G(z)} \sum_{k=n}^{\infty} k \binom{k-1}{n-1} \left[(1-p)\, G(z)\right]^k \\ &= \frac{n\, p^n \, g(z)\, G(z)^{n-1}}{\left[1-(1-p)\, G(z)\right]^{n+1}}. \end{aligned} \tag{5}$$

The cdf of Z (which holds for any positive real n) is given by

$$F(z) = \left[\frac{pG(z)}{1-(1-p)G(z)}\right]^n. \tag{6}$$

Inserting Eqs. (1) and (2) in Eq. (5), we obtain

$$f(z) = \frac{2\pi^{-1}\, n p^n \, (\alpha/\sigma)\, (z/\sigma)^{\alpha-1} \left[1+(z/\sigma)^{2\alpha}\right]^{-1} \left[2\pi^{-1} \tan^{-1}(z/\sigma)^{\alpha}\right]^{n-1}}{\left[1-2\pi^{-1}(1-p)\tan^{-1}(z/\sigma)^{\alpha}\right]^{n+1}}, \tag{7}$$

where $\alpha, \sigma > 0$ and $p \in (0,1)$. Henceforth, we denote by $Z \sim \text{PCNB}(n, p, \alpha, \sigma)$ a random variable having the density (7). The cdf of Z is given by

$$F(z) = F(z; n, p, \alpha, \sigma) = \left[\frac{2\pi^{-1} p \tan^{-1}(z/\sigma)^{\alpha}}{1-2\pi^{-1}(1-p)\tan^{-1}(z/\sigma)^{\alpha}}\right]^n. \tag{8}$$

Clearly, if $p = n = 1$, the PCNB model is identical to the PC distribution (2). Moreover, the PCNB(n, p, α, σ) model has the following six sub-models:

(i) If $p = n = \alpha = 1$, it gives the half-Cauchy (HC) distribution;
(ii) If $\alpha = 1$, it reduces to the half-Cauchy negative binomial (HCNB) distribution;
(iii) If $n = 1$, it gives the PC-geometric distribution;
(iv) If $\alpha = n = 1$, it becomes the HC-geometric distribution;
(v) If $p = 1$, it reduces to the exponentiated-PC distribution;
(vi) If $p = n = 1$, it becomes the PC distribution.

Note that the special models given in (ii), (iii) and (iv) do not exist in the literature.

The survival function (sf), hrf and reversed hazard rate function (rhrf) of Z are given by $S(z) = 1 - F(z)$, $h(z) = f(z)/S(z)$ and $r(z) = f(z)/F(z)$, respectively, where $F(z)$ and $f(z)$ are defined before. Figures 1 and 2 display some plots of the density and hrf of Z for $\sigma = 1$ and different values of α, p and n. The plots in Fig. 1a and b reveal that the PCNB density can have different shapes such as right-skewed and reversed-J. The plots in Fig. 2a

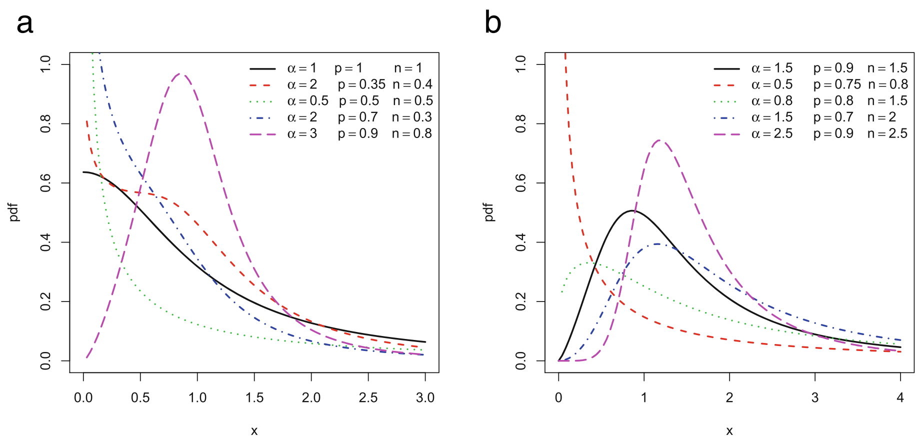

Fig. 1 Plots of the PCNB density

and b indicate that the hrf of Z can have DFR (decreasing failure rate), UBT and DID (decreasing-increasing-decreasing) shapes.

Properties of the new model

In this section, we provide some structural properties of the new distribution.

Quantile function and random number generation

The qf of Z is determined by inverting (8) as

$$Q(u) = \sigma \left[\tan \left(\frac{\pi\, u^{1/n}}{2\left[p + (1-p)\, u^{1/n}\right]} \right) \right]^{1/\alpha}, \quad u \in (0,1). \tag{9}$$

We can easily generate PCNB random variables from (9).

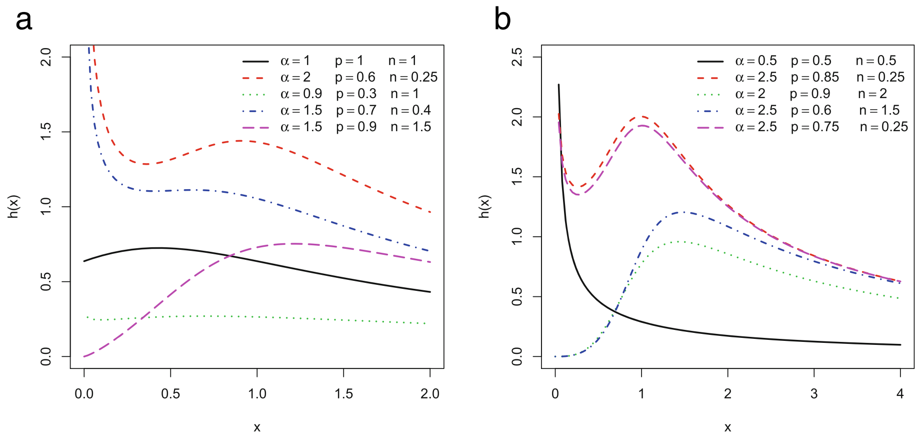

Fig. 2 Plots of the PCNB hrf

Tail behaviors

The tail behaviors of the pdf and cdf of Z in (7) and (8) are given as follow:

$$\begin{aligned} f(z) &\sim n\alpha A\, z^{n\alpha-1}, \text{as} \quad z \to 0^+, \\ f(z) &\sim \alpha B\, z^{-\alpha-1}, \text{as} \quad z \to \infty, \\ F(z) &\sim A\, z^{n\alpha}, \text{as} \quad z \to 0^+, \\ 1-F(z) &\sim B\, z^{-\alpha}, \text{as} \quad z \to \infty, \end{aligned}$$

where $A = (2p/\pi)^n$ and $B = 2n/(p\pi)$. For example, for fixed values of n and p, the left and right tails of the PCNB distribution are heavier when α increases. Also, for fixed values of α and p, the left tail becomes heavier when n increases.

Moments

For any real $q > 0$, the power series $(1-t)^{-q} = \sum_{j=0}^{\infty}(q)_j\, t^j/j!$ holds, where $(q)_j = q+(q+1)+\cdots+(q+j-1) = \Gamma(q+j)/\Gamma(q)$ is the ascending order factorial and $(q)_0 = 1$. Then, the cdf of Z in Eq. (8) can be expressed as

$$F(z) = \sum_{j=0}^{\infty} B_j(n,p)\, G_{PC}(z;\alpha,\sigma)^{n+j}, \tag{10}$$

where $B_j(n,p) = (n)_j\, p^n (1-p)^j/j!$ (for $j \geq 0$) and $G_{PC}(z;\alpha,\sigma)$ is the cdf given in Eq. (1).

By differentiating Eq. (10), the pdf of Z follows as

$$f(z) = \sum_{j=0}^{\infty} B_j(n,p)\, h_{n+j}(z), \tag{11}$$

where $h_{n+j}(z) = h_{n+j}(z;\alpha,\sigma)$ is the EPC density function with power parameter $n+j$ given by Eq. (4). Equation (11) reveals that the PCNB density is a linear combination of EPC densities. So, some mathematical properties of Z can be obtained from those of the EPC distribution. Next, we provide two examples.

Tahir et al. (2016) (see Sections 6.8 and 6.9) determined the sth ordinary and incomplete moments of Z_a as

$$\mathbf{E}\left(Z_a^s\right) = a\,\sigma^s \sum_{i=0}^{\infty} \frac{(0.5\,\pi)^{2i+\frac{s}{\alpha}}\; a_i(s/\alpha)}{(a+2\,i\,+s/\alpha)} \tag{12}$$

and

$$\int_0^z z^s f_{EPC}(z)dz = a\,\sigma^s \sum_{i=0}^{\infty} \frac{(0.5\,\pi)^{2i+s/\alpha}\, a_i(s/\alpha)\, D_z^{a+2i+s/\alpha}}{a+2\,i\,+s/\alpha}, \tag{13}$$

respectively, where $a_0(s) = 1$, $a_1(s) = s/3$, $a_2(s) = s(5s+7)/90$, etc, and $D_z = 2\,\pi^{-1}\,\tan^{-1}(z/\sigma)^{\alpha}$.

Then, the rth ordinary moment of Z follows from Eqs. (11) and (12) as

$$\mu'_r = \mathbb{E}(Z^r) = \sum_{i,j=0}^{\infty} B_j(n,p)\, \frac{(n+j)\,\sigma^r\;(0.5\,\pi)^{2i+r/\alpha}\;a_i(r/\alpha)}{\left(n+j\,+\,2\,i\,+\,r/\alpha\right)}. \tag{14}$$

Analogously, the rth incomplete moment of Z, say $m_r(z) = \int_0^z z^r f_{PCNB}(z)dz$, can be obtained from (11) and (13) as

$$m_r(z) = \sigma^r \sum_{i,j=0}^{\infty} B_j(n,p)\, a_i(r/\alpha)\, \frac{(n+j)\,(0.5\pi)^{2i+r/\alpha}\, D_z^{n+j+2i+r/\alpha}}{(n+j+2i+r/\alpha)}. \tag{15}$$

The first incomplete moment $m_1(q)$ follows from Eq. (15) for $r = 1$. It is useful to obtain the Bonferroni and Lorenz curves and mean deviations for the new model.

Estimation

Several approaches for parameter point estimation were proposed in the literature but the maximum likelihood method is the most commonly employed. The maximum likelihood estimates (MLEs) enjoy desirable properties that can be used when constructing confidence intervals for the model parameters. Large sample theory for these estimates delivers simple approximations that work well in finite samples. The normal approximation for the MLEs in distribution theory is easily handled either analytically or numerically.

We consider the estimation of the unknown parameters of the new distribution by the maximum likelihood method. Let $z_1, \ldots, z_m$ be m observed values from the PCNB distribution given by (7) with vector of parameters $\boldsymbol{\theta} = (n, p, \alpha, \sigma)^\top$. The log-likelihood $\ell = \ell(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ is given by

$$\begin{aligned}\ell = {} & m \log\left[2\, n\, p^n\, \pi^{-1}\, (\alpha/\sigma)\right] + (\alpha - 1)\sum_{i=1}^{m} \log\left(z_i/\sigma\right) - \sum_{i=1}^{m} \log\left[1 + (z_i/\sigma)^{2\alpha}\right] \\ & + (n-1)\sum_{i=1}^{m} \log[\, (2/\pi)\tan^{-1}(z_i/\sigma)^{\alpha}\,] \\ & - (n+1)\sum_{i=1}^{m} \log\left\{1 - (1-p)[\, (2/\pi)\tan^{-1}(z_i/\sigma)^{\alpha}\,]\right\}. \qquad (16)\end{aligned}$$

Equation (16) can be maximized either directly by using well-known computing platforms such as the R (`optim` function), SAS (`PROC NLMIXED`) and Ox program (subroutine `MaxBFGS`). These scripts can be applied and executed for a wide range of initial values. This process often leads to more than one maximum. However, in these cases, we consider the MLEs corresponding to the largest value of the log-likelihood statistics. In a few cases, no maximum is identified for the selected initial values. In these cases, new initial values can be tried in order to obtain a maximum. There exist sufficient conditions for the existence of the MLEs such as compactness of the parameter space and the concavity of the log-likelihood function. These estimates can exist even when such conditions are not satisfied. For more complex models, and in particular when there is no explicit solution, it is nearly impossible to establish theoretical conditions on the existence and uniqueness of the MLEs. However, such properties can be investigated numerically for this distribution and a given data set.

For interval estimation on the model parameters, we can evaluate the estimated observed information matrix $J(\widehat{\boldsymbol{\theta}})$ numerically. Further, we can easily check if the fit using the PCNB model is statistically "superior" to the fits using any of its six special models. For example, for comparing the PCNB and HC distributions, i.e., testing the null hypothesis $H_0 : p = n = \alpha = 1$ against $H_1 : H_0$ is false, the likelihood ratio (LR) statistic is given by $w = 2\{\ell(\widehat{\theta}) - \ell(\widetilde{\theta})\}$, where $\widehat{\theta}$ and $\widetilde{\theta}$ are the unrestricted and restricted estimates obtained by maximizing $\ell = \ell(\theta)$ under H_1 and H_0, respectively. The limiting distribution of this statistic is χ_3^2 under the null hypothesis, which is rejected if w exceeds the upper $100(1-\gamma)\%$ quantile of the χ_3^2 distribution.

The PCNB survival function has closed-form expression and hence this distribution can be used effectively in analyzing lifetime data in the presence of censoring. Consider a situation, where the time to event is not completely observed and is subjected to right censoring. Let C_i denote censoring time. We then observe $z_i = \min(t_i, c_i)$, where t_i is the observed time to the event and c_i is the observed right-censored, for $i = 1, \ldots, m$. The log-likelihood function reduces to

$$\begin{aligned}\ell(\boldsymbol{\theta}) = c_i \sum_{i=1}^{m} \Big\{ & \log\left[2np^n\pi^{-1}(\alpha/\sigma)\right] + (\alpha-1)\log(z_i/\sigma) \\ & - \log\left[1 + (z_i/\sigma)^{2\alpha}\right] - (m-1)\log\left[2\pi^{-1}\tan^{-1}(z_i/\sigma)^{\alpha}\right] \\ & -(m+1)\log\left[1 - 2(1-p)\pi^{-1}\tan^{-1}(z_i/\sigma)^{\alpha}\right]\Big\} \\ & +(1-c_i)\sum_{i=1}^{m}\log\left\{1 - \left[\frac{2p\pi^{-1}\tan^{-1}(z_i/\sigma)^{\alpha}}{1-2(1-p)\pi^{-1}\tan^{-1}(z_i/\sigma)^{\alpha}}\right]\right\}.\end{aligned}$$

The above log-likelihood can be maximized numerically to obtain the MLEs. We use the `optim` routine in the `R` software.

Monte Carlo simulation study. Now we assess the performance of the maximum likelihood method for estimating the PCNB parameters using Monte Carlo simulations. The simulation study is repeated 5000 times each with sample sizes $m = 50, 100, 200, 500$ and parameter scenarios: I: $p = 0.8$, $n = 0.5$, $\alpha = 0.5$ and $\sigma = 1$, II: $p = 0.5$, $n = 0.5$, $\alpha = 1.5$ and $\sigma = 1$ and III: $p = 0.1$, $n = 1.5$, $\alpha = 1.5$ and $\sigma = 1$. Table 1 gives the average biases (Bias) of the MLEs, mean square errors (MSE) and model-based coverage probabilities (CP) for the parameters p, n, α and σ under these scenarios and different sample sizes. Based on the simulation results, we conclude that the MLEs perform well in estimating the parameters of the PCNB distribution. The CPs of the confidence intervals are quite close to the 95% nominal levels. Therefore, the MLEs and their asymptotic results can be adopted for estimating and constructing confidence intervals for the model parameters.

Table 1 Monte Carlo simulation results: Biases, MSEs and CPs

		I			II			III		
Parameter	m	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP
p	50	− 0.233	0.218	0.94	− 0.021	0.169	0.94	0.205	0.195	0.93
	100	− 0.209	0.200	0.96	0.016	0.154	0.98	0.154	0.147	0.96
	200	− 0.188	0.183	0.96	0.055	0.150	0.96	0.111	0.100	0.95
	500	− 0.142	0.143	0.95	0.054	0.127	0.96	0.092	0.080	0.95
n	50	− 0.038	0.045	0.96	0.076	0.172	0.98	1.104	0.896	0.95
	100	− 0.036	0.024	0.98	− 0.016	0.022	0.96	0.983	0.521	0.95
	200	− 0.034	0.012	0.95	− 0.007	0.014	0.95	0.514	0.165	0.96
	500	− 0.027	0.006	0.95	− 0.010	0.006	0.95	− 0.024	0.013	0.95
α	50	0.124	0.071	0.99	0.370	0.217	0.98	0.077	0.153	0.95
	100	0.072	0.025	0.98	0.172	0.162	0.98	0.005	0.062	0.97
	200	0.046	0.010	0.96	0.093	0.074	0.97	− 0.012	0.031	0.95
	500	0.028	0.004	0.95	0.061	0.034	0.95	− 0.024	0.013	0.95
σ	50	0.312	0.573	0.93	− 0.033	0.351	0.92	0.069	1.010	1.00
	100	− 0.042	0.290	0.95	− 0.025	0.219	0.96	0.083	1.000	1.00
	200	− 0.096	0.107	0.96	− 0.018	0.168	0.96	0.101	0.902	0.99
	500	− 0.107	0.098	0.95	− 0.004	0.113	0.98	0.100	0.725	0.95

Regression model

In many practical applications, the lifetimes are affected by explanatory variables such as the cholesterol level, blood pressure, weight and many others. Parametric models to estimate univariate survival functions and for censored data regression problems are widely used. A regression model that provides a good fit to lifetime data tends to yield more precise estimates of the quantities of interest.

In applications in the area of survival analysis, the hrf is often U-shaped or unimodal, i.e., the function is not monotonic. The regression models commonly used for survival data are the log-Weibull, monotonic failure rate, log-logistic, decreasing failure rate and unimodal functions. One of the objectives of this work is to propose a new regression model, in location and scale form, called the *log-power-Cauchy negative-binomial* (LPCNB) regression model, which presents different failure rate functional forms. The proposed model is an alternative to the traditional extreme value (or log-Weibull), logistic and log-normal models, among others. One way to study the effect of these explanatory variables on the response variable Y is through a location-scale regression model, also known as a model of accelerated lifetime. These models consider that the response variable belongs to a family of distributions characterized by a location parameter and a scale parameter. Further details on this class of regression models can be found in Cox and Oakes (1984), Kalbfleisch and Prentice (2002) and Lawless (2003). In the context of survival analysis, some distributions have been used to analyze censored data. For example, more recently, Cruz et al. (2016) proposed the log-odd log-logistic Weibull regression model with censored data, Lanjoni et al. (2016) defined an extended Burr XII regression model and Ortega et al. (2016) introduced the odd Birnbaum-Saunders regression model with applications to lifetime data. In a similar manner, we define a location-scale regression model using the LPCNB regression model.

Let $Z \sim \text{PCNB}(n,p,\alpha,\sigma)$ be a random variable having the density (7). A class of regression models for location and scale is characterized by the fact that the random variable $Y = \log(Z)$ has a distribution with location parameter $\mu(\mathbf{v})$, which depends only on the explanatory variable vector, and a scale parameter a. Then, we can write $Y = \mu(\mathbf{v}) + aW$, where $a > 0$ and the distribution of W does not depend on $\mathbf{v}$.

The random variable $Y = \log(X)$ re-parameterized in terms of $\mu = \log(\sigma)$ and $a = \alpha^{-1}$ has density function (for $y \in \mathbb{R}$) given by

$$f(y) = \left(\frac{2p}{\pi}\right)^n \frac{n \exp\left(\frac{y-\mu}{a}\right) \arctan^{(n-1)}\left[\exp\left(\frac{y-\mu}{a}\right)\right]}{a\left\{1-(1-p)\,2\,\pi^{-1}\arctan\left[\exp\left(\frac{y-\mu}{a}\right)\right]\right\}^{(n+1)}}, \tag{17}$$

where $n > 0$ and $p \in (0,1)$ are shape parameters, $\mu \in \mathbb{R}$ is the location parameter and $a > 0$ is the scale parameter.

We refer to Eq. (17) as the LPCNB distribution, say $Y \sim \text{LPCNB}(n,p,\mu,a)$. If $Z \sim \text{PCNB}(n,p,\alpha,\sigma)$, then $Y = \log(Z) \sim \text{LPCNB}(n,p,\mu,a)$.

For $p = n = 1$, we obtain the log-power Cauchy (LPC) model. The survival function corresponding to Eq. (17) is given by

$$S(y) = 1 - \left(\frac{2p}{\pi}\right)^n \frac{\arctan^{n}\left[\exp\left(\frac{y-\mu}{a}\right)\right]}{\left\{1-(1-p)\,2\,\pi^{-1}\arctan\left[\exp\left(\frac{y-\mu}{a}\right)\right]\right\}^{n}}. \tag{18}$$

Plots of the density function (17) for selected parameter values are displayed in Fig. 3a and b, which show great flexibility for different values of p and n.

We define the standardized random variable $W = (Y - \mu)/a$ having the density function

$$f(w) = \left(\frac{2\,p}{\pi}\right)^{n} \frac{n\,\exp(w)\arctan\left[\exp(w)\right]^{(n-1)}}{\left\{1-(1-p)\,2\,\pi^{-1}\arctan\left[\exp(w)\right]\right\}^{(n+1)}}. \tag{19}$$

Next, we propose a linear location-scale regression model linking the response variable y_i and the explanatory variable vector $\mathbf{v}_i^T = (v_{i1}, \ldots, v_{ip})$ given by

$$y_i = \mathbf{v}_i^T \boldsymbol{\tau} + a\,w_i,\;\; i = 1, \ldots, m, \tag{20}$$

where the random error w_i has density function (19), $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^T$, $a > 0$, $n > 0$ and $p \in (0, 1)$ are unknown parameters. The parameter $\phi_i = \mathbf{v}_i^T \tau$ is the location of y_i. The location parameter vector $\phi = (\phi_1, \ldots, \phi_m)^T$ is represented by a linear model $\phi = \mathbf{v}\tau$, where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)^T$ is a known model matrix. The LPCNB model (20) opens new possibilities for fitting many different types of data.

Consider a sample $(y_1, \mathbf{v}_1), \ldots, (y_m, \mathbf{v}_m)$ of m independent observations, where each random response is defined by $y_i = \min\{\log(z_i), \log(c_i)\}$. We assume non-informative censoring such that the observed lifetimes and censoring times are independent. Let F and C be the sets of individuals for which y_i is the log-lifetime or log-censoring, respectively. Conventional likelihood estimation techniques can be applied here. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} = \left(p, n, a, \tau^T\right)^T$ from model (20) has the form $l(\boldsymbol{\theta}) = \sum\limits_{i \in F} l_i(\boldsymbol{\theta}) + \sum\limits_{i \in C} l_i^{(c)}(\boldsymbol{\theta})$, where $l_i(\boldsymbol{\theta}) = \log[f(y_i)]$, $l_i^{(c)}(\boldsymbol{\theta}) = \log[S(y_i)]$, $f(y_i)$ is the density (17) and $S(y_i)$ is the survival function (18) of Y_i. The total log-likelihood function for $\boldsymbol{\theta}$ reduces to

$$\begin{aligned} l(\boldsymbol{\theta}) \;=\; & q\log\left(\frac{n\,p^n\,2^n}{a\,\pi^n}\right) + \sum_{i \in F} w_i + (n-1)\sum_{i \in F}\log\{\arctan[\,\exp(w_i)]\,\} - \\ & (n+1)\sum_{i \in F}\log\left\{1-(1-p)2\,\pi^{-1}\arctan[\,\exp(w_i)]\,\right\} + \\ & \sum_{i \in C}\log\left\{1-\left(\frac{2\,p}{\pi}\right)^{n}\frac{\arctan^{n}[\,\exp(z_i)]}{\left\{1-(1-p)2\,\pi^{-1}\arctan[\,\exp(z_i)]\,\right\}^{n}}\right\}, \end{aligned} \tag{21}$$

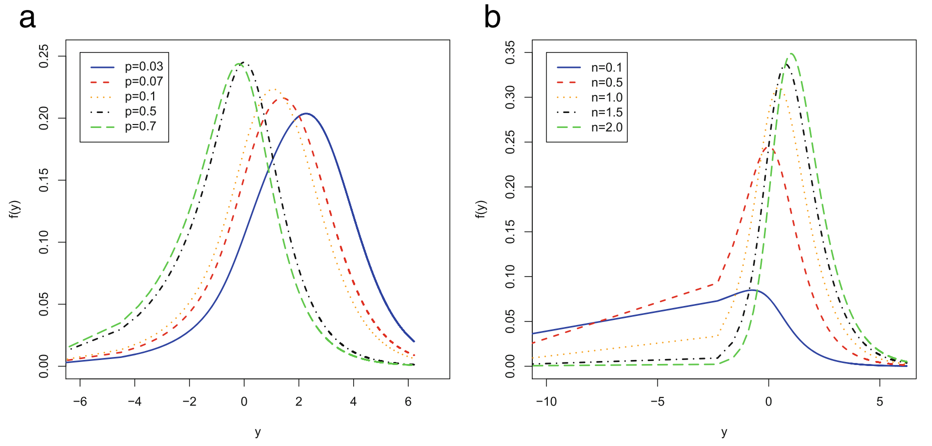

Fig. 3 Plots of the LPCNB density for some parameter values

where q is the number of uncensored observations (failures) and $w_i = \left(y_i - \mathbf{v}_i^T \boldsymbol{\tau}\right)/a$. The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be evaluated by maximizing the log-likelihood (21). We use the procedure NLMixed in SAS to calculate $\widehat{\boldsymbol{\theta}}$. Initial values for τ and a are taken from the fit of the LPC regression model with $p = n = 1$.

The elements of the $(p+3) \times (p+3)$ observed information matrix $J(\boldsymbol{\theta})$, namely $J_{pp}, J_{pn}, J_{pa}, J_{p\tau_j}, J_{nn}, J_{na}, J_{n\tau_j}, J_{aa}, J_{a\tau_j}$ and $J_{\tau_j\tau_s}$ (for $j, s = 1, \ldots, p$), can be evaluated numerically. Inference on $\boldsymbol{\theta}$ can be conducted in the classical way based on the approximate multivariate normal $N_{p+3}\left(0, J(\widehat{\boldsymbol{\theta}})^{-1}\right)$ distribution for $\widehat{\boldsymbol{\theta}}$.

We can use the likelihood ratio (LR) statistic for comparing some special models with the LPCNB regression model. We consider the partition $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T\right)^T$, where $\boldsymbol{\theta}_1$ is a subset of parameters of interest and $\boldsymbol{\theta}_2$ is a subset of remaining parameters. The LR statistic for testing the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^{(0)}$ versus the alternative hypothesis $H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^{(0)}$ is given by $w = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widetilde{\boldsymbol{\theta}})\}$, where $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$ are the estimates under the null and alternative hypotheses, respectively. The statistic w is asymptotically (as $n \to \infty$) distributed as χ_q^2, where q is the dimension of the subset of parameters $\boldsymbol{\theta}_1$ of interest.

Applications

In this section, the PCNB distribution is fitted to model three real life data sets. We compare the fits of the PCNB model with the beta-Weibull (BW) proposed by Lee et al. (2007), beta half-Cauchy (BHC) defined by Cordeiro and Lemonte (2011), Kumaraswamy half-Cauchy (KHC) presented by Ghosh (2014), power-Cauchy geometric (PCG) and power-Cauchy models. We estimate the parameters by using the maximum likelihood method. In order to compare the models, we consider the following goodness-of-fit statistics: Akaike information criterion (AIC) and Kolmogorov-Smirnov (K-S) measure with the associated p-value. The pdfs of the BW, BHC and KHC (for $x > 0$ and $a, b, c, \sigma, \lambda > 0$) distributions are given by

$$
\begin{aligned}
f_{BW}(x) &= \frac{c}{\lambda\, B(a,b)} \left(\frac{x}{\lambda}\right)^{c-1} \mathrm{e}^{-b(\frac{x}{\lambda})^c} \left[1 - \mathrm{e}^{-(\frac{x}{\lambda})^c}\right]^{a-1}, \\
f_{BHC}(x) &= K_1 \left[1 + \left(\frac{x}{\sigma}\right)^2\right]^{-1} \left[\tan^{-1}\left(\frac{x}{\sigma}\right)\right]^{a-1} \left[1 - 2\pi^{-1} \tan^{-1}\left(\frac{x}{\sigma}\right)\right]^{b-1}, \\
f_{KHC}(x) &= K_2 \left[1 + \left(\frac{x}{\sigma}\right)^2\right]^{-1} \left[\tan^{-1}\left(\frac{x}{\sigma}\right)\right]^{a-1} \left[1 - \left\{2\pi^{-1} \tan^{-1}\left(\frac{x}{\sigma}\right)\right\}^{a}\right]^{b-1},
\end{aligned}
$$

respectively, where $K_1 = 2^a/[\sigma\, \pi^a\, B(a,b)]$ and $K_2 = a\, b\, 2^a/(\sigma\, \pi^a)$.

Data set 1: Load haul dumbp machines failure data. First, we consider data on the times between successive failures (TBFs) of load haul dumbp machines. The operation and maintenance cards of a fleet of 19 LHD machines were collected for a period of one year. These cards record times to failure, the engine clock hour and the reported failures in case of operation cards, and the times to repairs and actual repairs performed in case of maintenance cards (see Kumar et al. [1989, Appendix 2, Table B1]). The summary statistics of the data are: m= 50, $\bar{x}$= 45.88, s=51.76936, skewness=2.07528 and kurtosis=6.06486. The MLEs (with SEs in parentheses), the AIC and K-S statistics and their p-values are listed in Table 2. The figures in this table indicate that the PCNB model provides the best fit to the current data. Next, we provide the scaled TTT plot, see Aarset (1987), for

Table 2 MLEs, their SEs (in parentheses) and goodness-of-fit measures for the first data set

Distribution	Estimates				AIC	K-S	*p*-value
PCNB(α, p, n, σ)	1.8529	0.0029	0.3905	3.5171	487.1582	0.0791	0.9131
	(0.3737)	(0.0015)	(0.1375)	(1.6477)			
BW(a, b, c, λ)	8.9783	0.10422	0.5264	0.3086	489.5797	0.1028	0.6655
	(3.9535)	(0.0224)	(0.0250)	(0.0033)			
PCG(α, p, σ)	1.182	71.1917	0.9153		492.4754	0.0911	0.8009
	(0.1415)	(6.7322)	(0.9334)				
BHC(a, b, σ)	1.5514	0.9514	11.1816		499.1124	0.1256	0.4096
	(0.6308)	(0.3152)	(9.3295)				
KHC(a, b, σ)	1.3321	0.9188	14.3689		497.3825	0.1528	0.1936
	(0.8090)	(0.3935)	(7.9620)				
PC(α, σ)	1.0127	25.1088			492.7543	0.0877	0.8360
	(0.1271)	(5.2807)					

these data in Fig. 4b. The summary statistics and Fig. 4a and b reveal that the first data set is right-skewed with DID failure rate shape. So, the PCNB has the ability to fit right-skewed data with DID failure rate shape. For a visual comparison, we provide PP plots of the fitted models to these data in Fig. 5. Clearly, the PCNB model provides a closer fit to these data.

Data set 2: Jet Airplanes failure data. The second data set is taken from Porchan (1963), which represents the failure times of air conditioning system of 720 jet airplanes. A set of the summary statistics of the data are: m=213, $\bar{x}$= 93.14085, s=106.7636, skewness=2.11185 and kurtosis=4.92499. The results of the fitted distributions are presented in Table 3. We conclude that the PCNB model provides the best fit with lowest values of the AIC and K-S statistics and largest p-value. The scaled TTT plot for the second data set in Fig. 6b gives an indication of a decreasing failure rate shape. The summary statistics and Fig. 6a and b reveal that the second data set is right-skewed with decreasing failure shape. So, the PCNB distribution can be used effectively to model these data. The PP plots in Fig. 7 also support the results of Table 3.

In conclusion, the PCNB model is certainly an appropriate model for fitting the first two data sets.

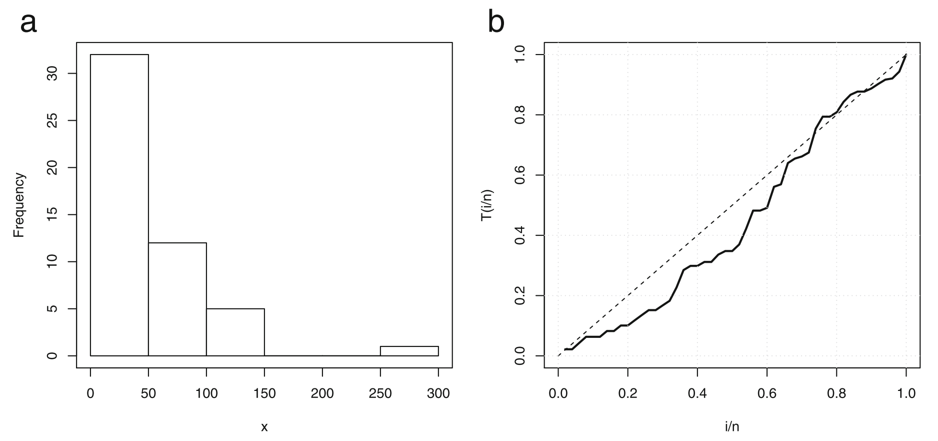

Fig. 4 **a** Histogram, **b** TTT plot for data set 1

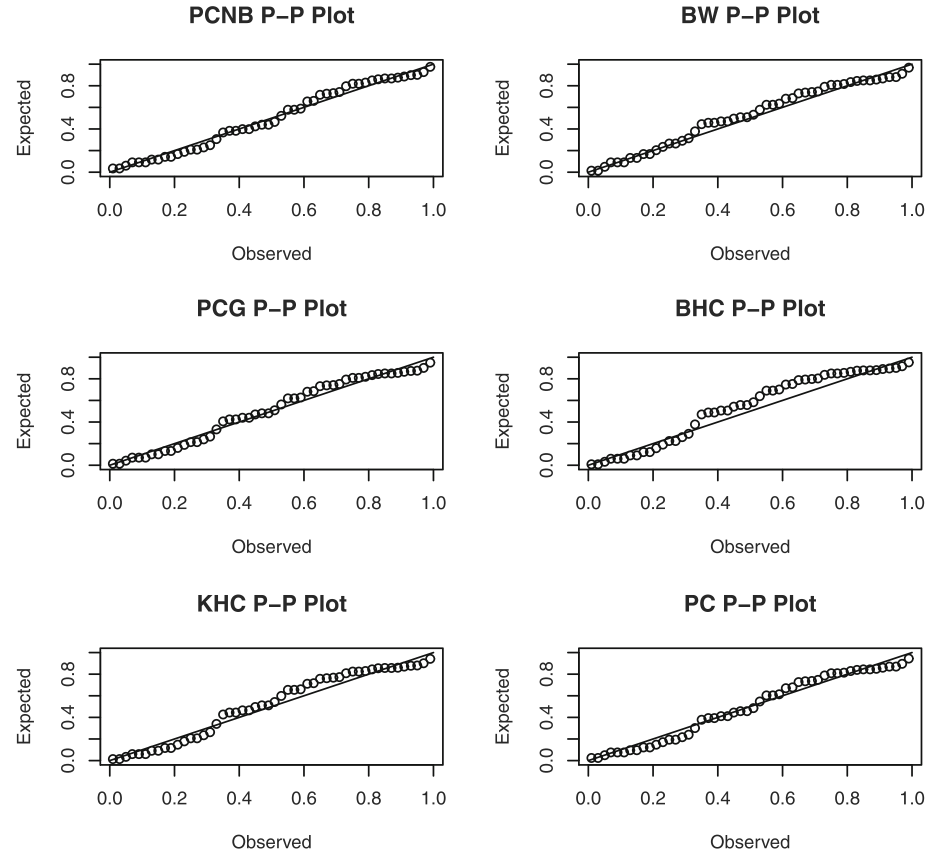

Fig. 5 P-P plots for data set 1

Data set 3: Head and neck cancer data. The third data set is taken from Efron (1988) regarding head and neck cancer clinical trial consisting of survival times of 51 patients in arm A who were given radiation therapy. Nine patients were lost to the follow-up and were regarded as censored observations. The MLEs of the model parameters are listed in Table 4. The figures in this table indicate that the PCNB model provides the best fit with

Table 3 MLEs, their SEs (in parentheses) and goodness-of-fit measures for the second data set

Distribution	Estimates				AIC	K-S	*p*-value
PCNB(α, p, n, σ)	1.6228	0.0035	0.7221	2.8888	2364.093	0.046	0.7588
	(0.1972)	(0.0009)	(0.1969)	(1.7041)			
BW(a, b, c, λ)	7.6164	0.1194	0.568	1.1134	2367.086	0.0767	0.1632
	(1.5171)	(0.0088)	0.0025)	(0.0033)			
PCG(α, p, σ)	1.3668	76.7098	2.9658		2368.893	0.0483	0.6507
	(0.0905)	(131.7613)	(3.3747)				
BHC(a, b, σ)	1.7824	0.9400	22.1049		2388.513	0.1029	0.0219
	(0.2498)	(0.1014)	(4.6448)				
KHC($a, b, \sigma a$)	1.4395	1.1559	36.2886		2375.904	0.0837	0.1013
	(0.1744)	(0.1462)	(7.4700)				
PC(α, σ)	1.1812	53.1034			2369.847	0.0524	0.6033
	(0.0723)	(4.5017)					

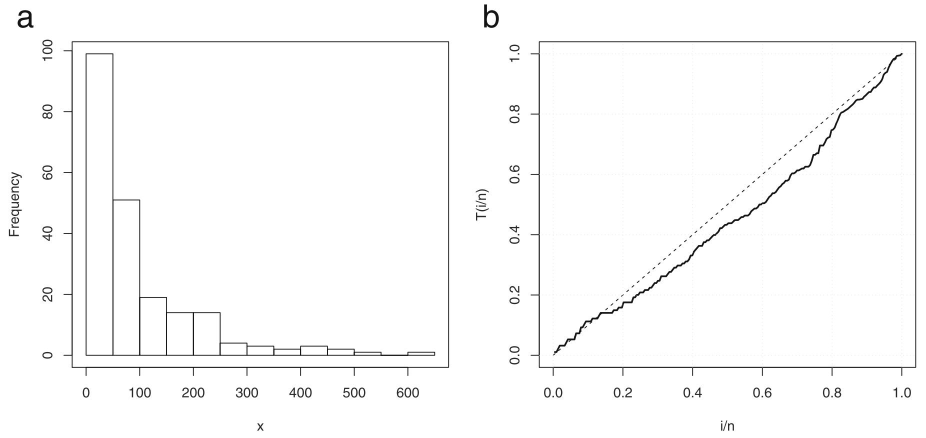

Fig. 6 **a** Histogram, **b** TTT plot for data set 2

lowest values of the AIC and K-S statistics. The plot in Fig. 8b reveals that the third data set has UBT failure rate shape, and then the PCNB distribution can be used effectively to model these data. The plots of the estimated survival functions of the PCNB, BW and GPC distributions are displayed in Fig. 8a. Clearly, the PCNB estimated survival function provides a closer fit to the empirical survival function than the other models.

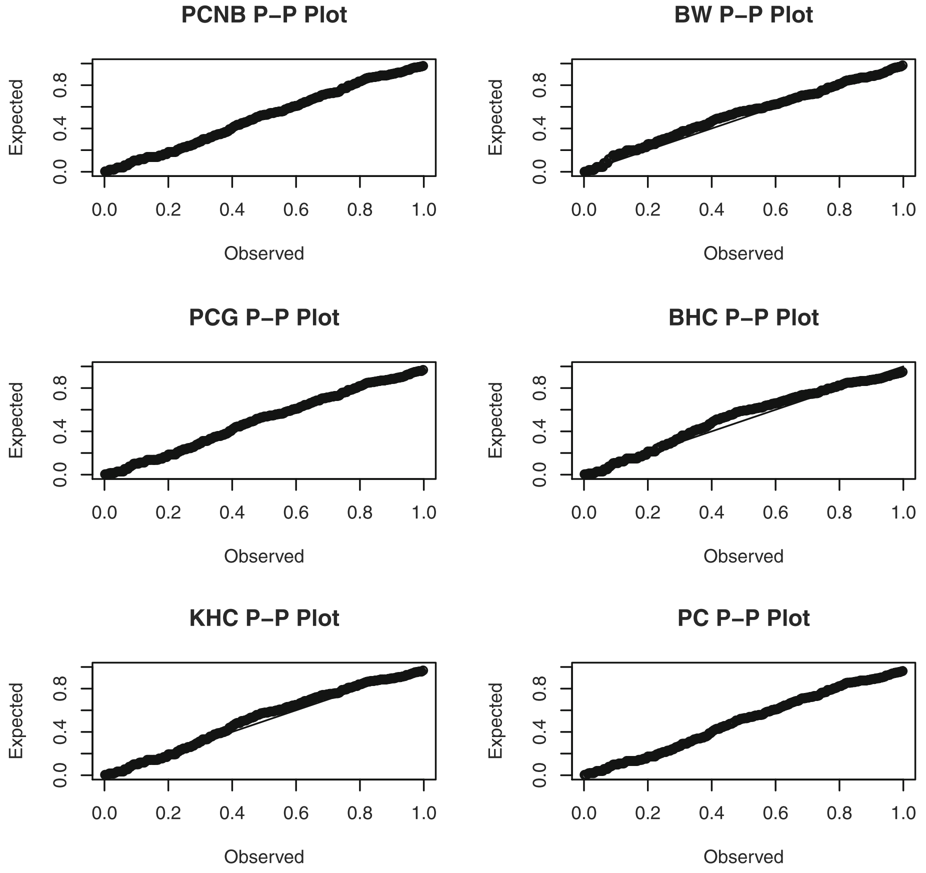

Fig. 7 P-P plots for data set 2

Table 4 MLEs, their SEs (in parentheses) and goodness-of-fit measures for the third data set

Distribution	Estimates				AIC	BIC
PCNB(α, p, n, σ)	1.2865	0.0070	1.6620	4.2398	592.1857	599.9130
	(0.2690)	(0.0055)	(1.0335)	(6.2151)		
BW(a, b, c, λ)	17.8517	0.7694	0.3143	4.0437	594.1023	601.8296
	(59.5446)	(1.8538)	(0.4824)	(25.6600)		
PCG(α, p, σ)	1.4738	0.0053	9.2046		599.0166	604.8121
	(0.1888)	(0.0041)	(6.0303)			
BHC(a, b, σ)	1.9480	1.0755	127.2517		598.7476	604.5431
	(0.6174)	(0.2655)	(54.0490)			
KHC(a, b, σ)	1.9059	1.0921	130.6862		598.7297	604.5252
	(0.6031)	(0.2780)	(55.0280)			
PC(α, σ)	0.5788	45.3099			649.3128	653.1764
	(0.1101)	(12.3464)				

Regression model example : Entomology data. First, we use the data from a study carried out at the Department of Entomology of the Luiz de Queiroz School of Agriculture, University of São Paulo, which aims to assess the longevity of the Mediterranean fruit fly (ceratitis capitata). The need for this fly to seek food just after emerging from the larval stage has permitted the use of toxic baits for its management in Brazilian orchards for at least fifty years. This pest control technique consists of using small portions of food laced with an insecticide, generally an organophosphate, that quickly kills the flies, instead of using an insecticide alone. Recently, there have been reports of the insecticidal effect of extracts of the neem tree leading to proposals to adopt various extracts (aqueous extract of the seeds, methanol extract of the leaves and dichloromethane extract of the branches) to control pests such as the Mediterranean fruit fly. The experiment was completely randomized with eleven treatments, consisting of different extracts of the neem tree, at concentrations of 39, 225 and 888 ppm.

After preliminary statistical analysis, these eleven treatments were allocated into two groups, namely:

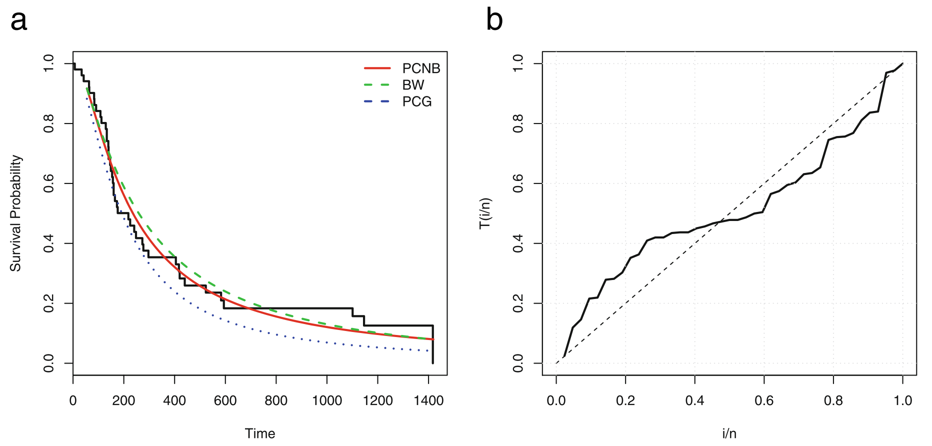

Fig. 8 **a** Empirical survival and estimated survival functions, **b** TTT plot for data set 3

Table 5 MLEs of the parameters from the LPCNB regression model fitted to the entomology data set, the corresponding SEs (given in parentheses), p-values in [·]

Model	a	n	p	τ_0	τ_1	τ_2
LPCNB	0.2514	0.4118	0.1496	2.9793	0.0188	-0.2787
	(0.0383)	(0.0897)	(0.1470)	(0.1471)	(0.0779)	(0.0854)
				[< 0.001]	[0.8098]	[0.0013]
LPC	0.4100	1	1	3.0781	-0.0207	-0.2779
	(0.0293)			(0.0617)	(0.0832)	(0.0939)
				[< 0.001]	[0.8038]	[0.0035]

- Group 1: Control 1 (deionized water); Control 2 (acetone - 5%); aqueous extract of seeds (AES) (39 ppm); AES (225 ppm); AES (888 ppm); methanol extract of leaves (MEL) (225 ppm); MEL (888 ppm); and dichloromethane extract of branches (DMB) (39 ppm).
- Group 2: MEL (39 ppm); DMB (225ppm) and DMB (888 ppm).

The response variable in the experiment is the lifetime of the adult flies in days after exposure to the treatments. The experimental period was set at 51 days, so that the numbers of larvae that survived beyond this period were considered as censored data. The total sample size is $n = 72$, because four observations were lost. Therefore, the variables used in this study are: z_i-lifetime of ceratitis capitata adults in days, v_{i1}-sex of the larvae and v_{i2}-group (0=group 1, 1=group 2). We start the analysis of these data considering only failure (z_i) and censoring (c_i) data and an appropriate model for fitting the data could be the LPCNB and LPC distributions.

Next, we present results on fitting the model

$$y_i = \tau_0 + \tau_1 v_{i1} + \tau_2 v_{i2} + a w_i,$$

where the response variable Y_i follows the LPCNB distribution given in (17), $i = 1, \ldots, 72$. Table 5 lists the MLEs and their standard errors in parentheses for two fitted regression models. The MLEs of the model parameters are evaluated using the NLMixed procedure in SAS. Iterative maximization of the logarithm of the likelihood function (21) starts with initial values for $\boldsymbol{\tau}$ and σ, which are taken from the fit of the LPC regression model.

We note from the fitted LPCNB regression model that v_2 is significant at 1% and that there is a significant difference between the groups 1 and 2 for the survival times. Table 6 gives a summary of the AIC, consistent Akaike information criterion (CAIC) and Bayesian information criterion (BIC) to compare the LPCNB and LPC regression models. The LPCNB regression model outperforms the LPC model irrespective of the criteria and then they can be used effectively in the analysis of these data.

Finally, we turn to a simplified model retaining only v_2 as an explanatory variable

$$y_i = \tau_0 + \tau_2 v_{i2} + a\, w_i.$$

Table 6 AIC, CAIC and BIC statistics for comparing the LPCNB and LPC regression models

Model	AIC	CAIC	BIC
LPCNB	332.8	333.4	351.7
LPC	346.0	346.0	358.6

Table 7 MLEs of the parameters from the fitted LPCNB regression model to the entomology data

Model	a	n	p	τ_0	τ_2
LPCNB	0.2532	0.4164	0.1569	2.9917	-0.2771
	(0.0375)	(0.0879)	(0.1496)	(0.1391)	(0.0846)
				[< 0.001]	[0.0013]

The MLEs for the LPCNB regression model fitted to these data are listed in Table 7. In order to assess if the model is appropriate, Fig. 9 displays the plots of the empirical survival function and the estimated survival function from the fitted LPCNB regression model. In fact, this regression model provides a good fit to these data.

Concluding remarks

We consider a lifetime model in the context of insurance claims where the claim sizes follow a power Cauchy and the number of claims is negative binomial distributed. In these terms, we propose a new model by compounding the power-Cauchy and negative-binomial distributions called the *power-Cauchy negative-binomial* (PCNB) distribution. We provide a useful linear representation for its density, which allows to obtain some properties for the proposed distribution. We use the maximum likelihood method for estimating the model parameters. The suitability of these estimates is investigated by a simulation study. We fit the proposed distribution to three real data sets to show empirically its flexibility. We proposed a new class of regression models for location and scale based on the logarithm of the PCNB random variable. Estimation and inference on the regression coefficients are discussed and an application to real data in Entomology is addressed. Various future studies can be conducted, such as employing other estimation techniques (bootstrap and Bayesian methods) and investigating the sensitivity of the LPCNB regression model using diagnosis and analysis of residuals. which led to this improved version.

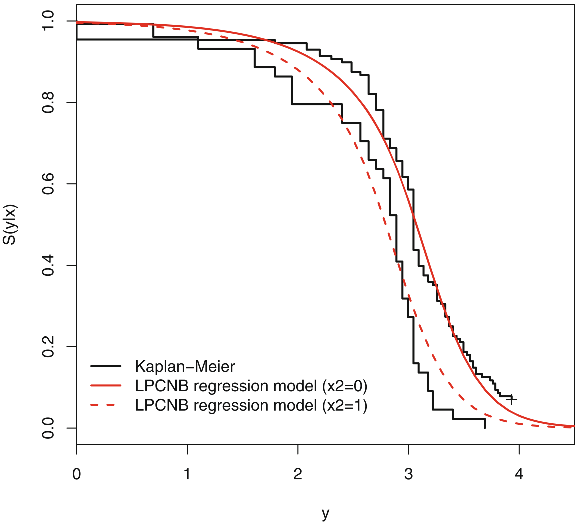

Fig. 9 Estimated LPCNB survival function and the empirical survival for entomology data

Acknowledgement
The authors are grateful to the Editor-in-Chief, the Associate Editor and anonymous referees for their many helpful comments and suggestions on an earlier version of this paper which led to this improved version.

Authors' contributions
The authors, viz MZ, MHT, GMC, AA and EMMO with the consultation of each other carried out this work and drafted the manuscript together. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Statistics, Govt. S.E. College, Bahawalpur, Pakistan, 63100 Bahawalpur, Pakistan. [2]Department of Statistics, The Islamia University of Bahawalpur, 63100 Bahawalpur, Pakistan. [3]Departamento de Estatística, Universidade Federal de Pernambuco, PE 50740-540, Recife, Brazil. [4]Department of Mathematics and Statistics, American University of Sharjah, 26666 Sharjah, UAE. [5]Departamento de Ciências Exatas, ESALQ, Universidade de São Paulo, Piracicaba/SP, Brazil.

Received: 24 January 2017 Accepted: 11 December 2017
Published online: 08 January 2018

References

Aarset, MV: How to identify bathtub hazard rate. IEEE Trans. Reliab. **36**, 106–108 (1987)
Cordeiro, GM, Lemonte, AJ: The beta-half Cauchy distribution. J. Probab. Statist. Art. ID. (904705), 18 (2011)
Cox, DR, Oakes, D: Analysis of survival data. Chapman and Hall, New York (1984)
Cruz, da, Ortega, JN, Cordeiro, EMM: GM: The log-odd log-logistic Weibull regression model: modelling, estimation, influence diagnostics and residual analysis. J. Stat. Comput. Simul. **86**, 1516–1538 (2016)
Efron, B: Logistic regression, survival analysis, and the Kaplan–Meier curve. J. Amer. Statist. Assoc. **83**, 414–425 (1988)
Ghosh, I: The Kumaraswamy-half Cauchy distribution: Properties and applications. J. Stat. Theory Appl. **13**, 122–134 (2014)
Kalbfleisch, JD, Prentice, RL: The statistical analysis of failure time data. Wiley, New York (2002)
Kumar, U, Klefsjo, B, Granholm, S: Reliability investigation for a fleet of load haul dump machines in a Swedish mine. Reliab. Eng. Syst. Safet. **26**, 341–361 (1989)
Lanjoni, BR, Ortega, EMM, Cordeiro, GM: Extended Burr XII regression models: Theory and applications. J. Agric. Biol. Environ. Stat. **21**, 203–224 (2016)
Lawless, JF: Statistical models and methods for lifetime data. Wiley, New Jersey (2003)
Lee, C, Famoye, F, Olumolade, O: Beta-Weibull distribution: Some properties and applications to censored data. J. Mod. Appl. Stat. Methods. **6**, 173–186 (2007)
Nekoukhou, V, Bidram, H: A new generalization of the Weibull-geometric distribution with bathtub failure rate. Commun. Stat. Theory Methods. **46**, 4296–4310 (2017)
Ortega, EMM, Lemonte, AJ, Cordeiro, GM, da Cruz, JN: The odd Birnbaum-Saunders regression model with applications to lifetime data. J. Stat. Theory Pract. **10**, 780–804 (2016)
Proschan, F: Theoretical explanation of observed decreasing failure rate. Technometrics. **5**, 375–383 (1963)
Rooks, B, Schumacher, A, Cooray, K: The power Cauchy distribution: derivation, description, and composite models. NSF-REU Program Reports (2010). Available from "http://www.cst.cmich.edu/mathematics/research/REU_and_LURE.shtml"
Tahir, MH, Zubair, M, Cordeiro, GM, Alzaatreh, A, Mansoor, M: The Poisson-X family of distributions. J. Stat. Comput. Simul. **86**, 2901–2921 (2016)