Statistical Modelling 2018; **18(1)**: 24–49

Extended Poisson–Tweedie: Properties and regression models for count data

Wagner H. Bonat[1,2], Bent Jørgensen[2], Célestin C. Kokonendji[3], John Hinde[4] and Clarice G. B. Demétrio[5]

[1]Laboratory of Statistics and Geoinformation, Department of Statistics, Paraná Federal University, Curitiba, Brazil.
[2]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark.
[3]Laboratoire de Mathématiques de Besançon, Bourgogne Franche-Comté University, Besançon, France.
[4]School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Galway, Ireland.
[5]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiroz, São Paulo University, Piracicaba, Brazil.

Abstract: We propose a new class of discrete generalized linear models based on the class of Poisson–Tweedie factorial dispersion models with variance of the form $\mu + \phi\mu^p$, where μ is the mean and ϕ and p are the dispersion and Tweedie power parameters, respectively. The models are fitted by using an estimating function approach obtained by combining the quasi-score and Pearson estimating functions for the estimation of the regression and dispersion parameters, respectively. This provides a flexible and efficient regression methodology for a comprehensive family of count models including Hermite, Neyman Type A, Pólya–Aeppli, negative binomial and Poisson-inverse Gaussian. The estimating function approach allows us to extend the Poisson–Tweedie distributions to deal with underdispersed count data by allowing negative values for the dispersion parameter ϕ. Furthermore, the Poisson–Tweedie family can automatically adapt to highly skewed count data with excessive zeros, without the need to introduce zero-inflated or hurdle components, by the simple estimation of the power parameter. Thus, the proposed models offer a unified framework to deal with under-, equi-, overdispersed, zero-inflated and heavy-tailed count data. The computational implementation of the proposed models is fast, relying only on a simple Newton scoring algorithm. Simulation studies showed that the estimating function approach provides unbiased and consistent estimators for both regression and dispersion parameters. We highlight the ability of the Poisson–Tweedie distributions to deal with count data through a consideration of dispersion, zero-inflated and heavy tail indices, and illustrate its application with four data analyses. We provide an `R` implementation and the datasets as supplementary materials.

Key words: count data, estimating functions, overdispersion, underdispersion, Poisson–Tweedie distribution

Received November 2016; **revised** April 2017; **accepted** May 2017

Address for correspondence: Wagner Hugo Bonat, Department of Statistics, Paraná Federal University, Centro Politécnico, Curitiba, 81531980 CP19081, Brazil.
E-mail: wbonat@ufpr.br

 10.1177/1471082X17715718

1 Introduction

Generalized linear models (GLMs) (Nelder and Wedderburn, 1972) have been the main statistical tool for regression modelling of normal and non-normal data over the past four decades. The success enjoyed by the GLM framework comes from its ability to deal with a wide range of normal and non-normal data. GLMs are fitted by a simple and efficient Newton-scoring algorithm, relying only on second-moment assumptions for estimation and inference. Furthermore, the theoretical background for GLMs is well established in the class of dispersion models (Jørgensen, 1987, 1997) as a generalization of the exponential family of distributions. In particular, the Tweedie family of distributions plays an important role in the context of GLMs, since it encompasses many special cases including the normal, Poisson, non-central gamma, gamma and inverse Gaussian.

In spite of the flexibility of the Tweedie family, the Poisson distribution is the only choice for the analysis of count data in the context of GLMs. For this reason, in practice, there is probably an over emphasis on the use of the Poisson distribution for count data. A well-known limitation of the Poisson distribution is its mean and variance relationship, which implies that the variance equals the mean, referred to as equidispersion. In practice, however, count data can present other features, namely underdispersion (mean > variance) and overdispersion (mean < variance) that is often related to zero inflation (ZI) or a heavy tail (HT). These departures can make the Poisson distribution unsuitable, or at least of limited use, for the analysis of count data. The use of the Poisson distribution for non-equidispersed data may cause problems, because, in case of overdispersion, standard errors (SEs) calculated under the Poisson assumption are too optimistic and associated hypothesis tests will tend to give false positive results by incorrectly rejecting null hypotheses. The opposite situation will appear in case of underdispersed data. In both cases, the Poisson model provides unreliable SEs for the regression coefficients and hence potentially misleading inferences.

The analysis of overdispersed count data has received much attention. Hinde and Demétrio (1998) discussed models and estimation algorithms for overdispersed data. Kokonendji et al. (2004, 2007) discussed the theoretical aspects of some discrete exponential models, in particular, the Hinde–Demétrio and Poisson–Tweedie classes. El-Shaarawi et al. (2011) applied the Poisson–Tweedie family for modelling species abundance. Rigby et al. (2008) presented a general framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. Rigby et al. (2008) also characterized many well-known distributions such as the negative binomial, Poisson-inverse Gaussian (PIG), Sichel, Delaporte and Poisson–Tweedie as Poisson mixtures. In general, these models are computationally slow to fit to large datasets, their probability mass functions cannot be expressed explicitly and they deal only with overdispersed count data.

The phenomenon of overdispersion is in general manifested through an HT and/or ZI. Zhu and Joe (2009) discussed the analysis of heavy-tailed count data based on the generalized PIG family, which corresponds to Poisson–Tweedie distributions with power parameter larger than 2. The problem of ZI has been well discussed

(Ridout et al., 1998) and solved by including hurdle or ZI components (Zeileis et al., 2008). These models are specified by two parts. The first part is a binary model for the dichotomous event of having zero or count values, for which the logistic model is a frequent choice. Conditional on a count value, the second part assumes a discrete distribution, such as the Poisson or negative binomial (Loeys et al., 2012), or zero-truncated versions for the hurdle model. While quite flexible, the two-part approach has the disadvantage of increasing the model complexity by having an additional linear predictor to describe the excess of zeros.

The phenomenon of underdispersion seems less frequent in practical data analysis; however, recently some authors have given attention towards the underdispersion phenomenon. Sellers and Shmueli (2010) presented a flexible regression model based on the Conway–Maxwell–Poisson (COM–Poisson) distribution that can deal with over- and underdispersed data. The COM–Poisson model has also recently been extended to deal with ZI (Sellers and Raim, 2016). Zeviani et al. (2014) discussed the analysis of underdispersed experimental data based on the Gamma-count distribution. Similarly, Kalktawi et al. (2015) proposed a discrete Weibull regression model to deal with under- and overdispersed count data. Although flexible, these approaches share the disadvantage that the probability mass function cannot be expressed explicitly, which implies that estimation and inference based on the likelihood function are difficult and time consuming. Furthermore, the expectation is not known in a closed form, which makes these distributions unsuitable for regression modelling, where, in general, we are interested in modelling the effects of covariates on a function of the expectation of the response variable.

Given the plethora of available approaches to deal with count data in the literature, it is difficult to decide, with conviction, which is the best approach for a particular dataset. The orthodox approach seems to be to take a small set of models, such as the Poisson, negative binomial, PIG, zero-inflated Poisson, zero-inflated negative binomial, etc., fit all of these models and then choose the best fit by using some measures of goodness of fit, such as the *Akaike* or Bayesian information criteria. A typical example of this approach can be found in Oliveira et al. (2016), where the authors compared the fit of eight different models for the analysis of datasets related to ionizing radiation. Although reasonable, such an approach is difficult to use in practical data analysis. The first problem is to define the set of models to be considered. Second, each count model can require specific fitting algorithms and give its own set of fitting problems, in general, due to the bad behaviour of the likelihood function. Third, the choice of the best fit may not be obvious, with different information criteria leading to different selected models. Finally, the uncertainty around the choice of distribution is not taken into account when choosing the best fit. Thus, we claim that it is very useful and attractive to have a unified model that can automatically adapt to the underlying dispersion and that can be easily implemented in practice.

The main goal of this article is to propose such a new class of count GLMs based on the class of Poisson–Tweedie factorial dispersion models (Jørgensen and Kokonendji, 2016) with variance of the form $\mu + \phi\mu^p$, where μ is the mean and ϕ and p are the dispersion and Tweedie power parameters, respectively. The proposed class provides

a unified framework to deal with over-, equi- or underdispersed, zero-inflated and heavy-tailed count data, with many potential applications.

As for GLMs, this new class relies only on second-moment assumptions, that is, expectation and variance for estimation and inference. Thus, our approach resembles Wedderburn's quasi-likelihood (Wedderburn, 1974) method and the generalized estimating equations of Liang and Zeger (1986) and Zeger et al. (1988), popular for the analysis of longitudinal data. In this framework, we do not specify a full probability mass distribution for the count response variable and consequently a likelihood function is not available. Thus, our models are fitted by an estimating function approach (Jørgensen and Knudsen, 2004; Bonat and Jørgensen, 2016), where the quasi-score and Pearson estimating functions are adopted for the estimation of regression and dispersion parameters, respectively. The estimating function combined with the second-moment assumptions allows us to extend the Poisson–Tweedie distributions to deal with underdispersed count data by allowing negative values for the dispersion parameter ϕ. The Tweedie power parameter plays an important role in the Poisson–Tweedie family since it is an index that distinguishes between important distributions, examples include Hermite ($p = 0$), Neyman Type A (NTA; $p = 1$), Pólya–Aeppli ($p = 1.5$), negative binomial (NB; $p = 2$) and PIG ($p = 3$). Furthermore, through the estimation of the Tweedie power parameter, the Poisson–Tweedie family automatically adapts to highly skewed count data with excessive zeros, without the need to introduce zero-inflated or hurdle components.

The Poisson–Tweedie family of distributions and its properties are introduced in Section 2. In Section 3, we considered the estimating function approach for parameter estimation and inference. Section 4 presents the main results of two simulation studies conducted to check the properties of the estimating function derived estimators and explore the flexibility of the extended Poisson–Tweedie models to deal with underdispersed count data. The application of extended Poisson–Tweedie regression models is illustrated in Section 5. Finally, discussions and directions for future work are given in Section 6. The R implementation and the datasets are available in the supplementary material.

2 Poisson–Tweedie: Properties and regression models

In this section, we derive the probability mass function and discuss the main properties of the Poisson–Tweedie distributions. Furthermore, we propose the extended Poisson–Tweedie regression model. The Poisson–Tweedie distributions are Poisson–Tweedie mixtures. Thus, our initial point is an exponential dispersion model of the form

$$f_Z(z; \mu, \phi, p) = a(z, \phi, p) \exp\{(z\psi - k_p(\psi))/\phi\},$$

where $\mu = \mathrm{E}(Z) = k_p'(\psi)$ is the mean, $\phi > 0$ is the dispersion parameter, ψ is the canonical parameter and $k_p(\psi)$ is the cumulant function. The variance is given by $\mathrm{Var}(Z) = \phi V(\mu)$, where $V(\mu) = k_p''(\psi)$ is called the variance function. Tweedie

densities are characterized by power variance functions of the form $V(\mu) = \mu^p$, where $p \in (-\infty, 0] \cup [1, \infty)$ is the index determining the distribution. For a Tweedie random variable Z, we write $Z \sim Tw_p(\mu, \phi)$. The support of the distribution depends on the value of the power parameter. For $p \geq 2$, $1 < p < 2$ and $p = 0$, the support corresponds to the positive, non-negative and real values, respectively. In these cases $\mu \in \Omega$, where Ω is the convex support (i.e., the interior of the closed convex hull of the corresponding distribution support). Finally, for $p < 0$, the support again corresponds to the real values; however, the expectation μ is positive.

The function $a(z, \phi, p)$ cannot be written in a closed form, apart from the special cases corresponding to the Gaussian ($p = 0$), Poisson ($\phi = 1$ and $p = 1$), non-central gamma ($p = 3/2$), gamma ($p = 2$) and inverse Gaussian ($p = 3$) distributions (Jørgensen, 1997; Bonat and Kokonendji, 2017). Another important case corresponds to the compound Poisson distributions, obtained when $1 < p < 2$. The compound Poisson distribution is a frequent choice for the modelling of non-negative data that has a probability mass at zero and is highly right-skewed (Smyth and Jørgensen, 2002).

The Poisson–Tweedie family is given by the following hierarchical specification:

$$\begin{aligned} Y|Z &\sim \text{Poisson}(Z) \\ Z &\sim \text{Tw}_p(\mu, \phi). \end{aligned}$$

Here, we require $p \geq 1$ to ensure that Z is non-negative. In this case, the Poisson–Tweedie is an overdispersed factorial dispersion model (Jørgensen and Kokonendji, 2016). The probability mass function for $p > 1$ is given by

$$f(y; \mu, \phi, p) = \int_0^\infty \frac{z^y \exp -z}{y!} a(z, \phi, p) \exp\{(z\psi - k_p(\psi))/\phi\} dz. \qquad (2.1)$$

The integral (2.1) has no simple form apart from the special case corresponding to the NB distribution, obtained when $p = 2$, that is, a Poisson gamma mixture. For $p = 1$, the integral (2.1) is replaced by a sum and we have the NTA distribution. Further, special cases include the Hermite ($p = 0$), Poisson compound Poisson ($1 < p < 2$), factorial discrete positive stable ($p > 2$) and PIG ($p = 3$) distributions (Jørgensen and Kokonendji, 2016; Kokonendji et al., 2004).

El-Shaarawi et al. (2011) using a slightly different parametrization proposed a recursive algorithm for computing the probability mass function of the Poisson–Tweedie distribution. This algorithm was employed by Esnaola et al. (2013) to achieve maximum likelihood estimation of Poisson–Tweedie regression models, and it is implemented in the package `tweeDEseq` for the statistical software `package R` (R Core Team, 2016). Recently, Barabesi et al. (2016) presented a finite sum expression for the probability mass function of the Poisson–Tweedie distribution.

Simulation from Poisson–Tweedie distributions is easy because of the availability of good simulation procedures for Tweedie distributions (Dunn, 2013). Although the aforementioned algorithms are available, we can approximate the integral (equation 2.1) using Monte Carlo integration, since the Tweedie family is a

natural proposal distribution. Alternatively, we can evaluate the integral using the Gauss–Laguerre method. Figure 1 presents the empirical probability mass function for some Poisson–Tweedie distributions computed based on a random sample of size 100 000 (grey). Additionally, we display an approximation for the probability mass function (black line) obtained by Monte Carlo integration. We considered different values of the Tweedie power parameter ($p = 1.1, 2, 3$) combined with different values of the dispersion index (DI $= 2,\ 5,\ 10,\ 20$), which is defined by

$$\mathrm{DI} = \mathrm{Var}(Y)/\mathrm{E}(Y).$$

In all scenarios, the expectation μ was fixed at 10.

Figure 1 shows that in the small dispersion case (DI $= 2$), the shape of the probability mass functions is quite similar for the different values of the power parameter. However, when the dispersion index increases, the differences become more marked. For $p = 1.1$, the overdispersion is clearly attributable to ZI, while for $p = 3$, the overdispersion is due to the HT. The NB case ($p = 2$) is a critical point, where the distribution changes from zero-inflated to heavy-tailed. The results in Figure 1 also show that the Monte Carlo method provides a reasonable approximation for the probability mass function for all Poisson–Tweedie distributions.

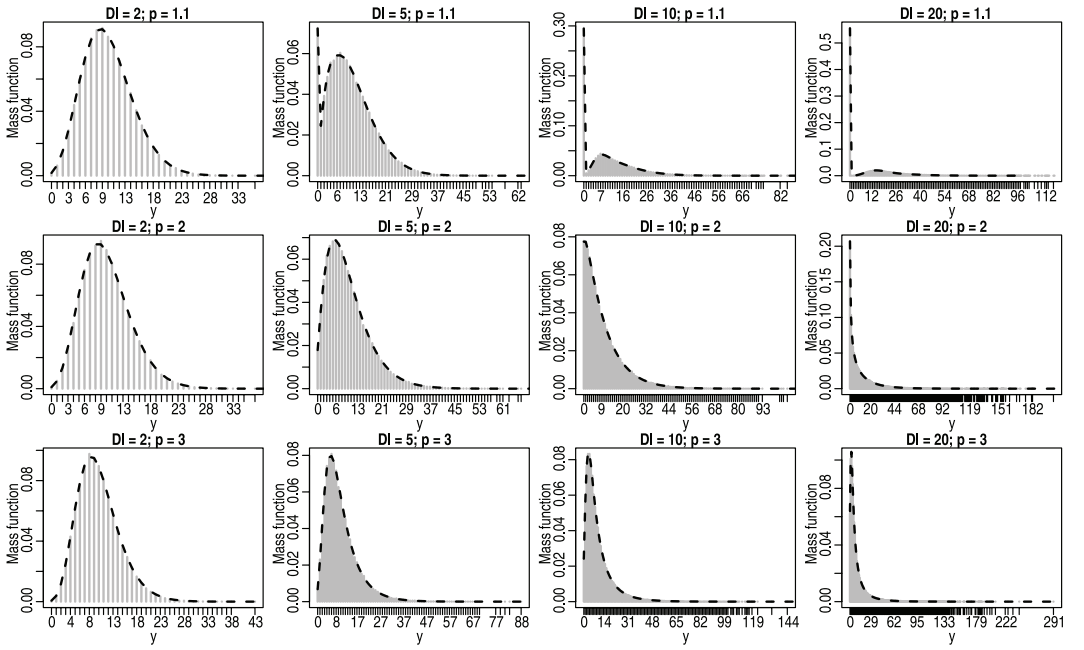

Figure 1 Empirical (grey) and approximated (black) Poisson–Tweedie probability mass function by values of the dispersion index (DI) and Tweedie power parameter

In order to further explore the flexibility of the Poisson–Tweedie distributions, we introduce indices for ZI

$$\text{ZI} = 1 + \frac{\log \text{P}(Y = 0)}{\text{E}(Y)}$$

and an HT

$$\text{HT} = \frac{\text{P}(Y = y + 1)}{\text{P}(Y = y)} \quad \text{for} \quad y \to \infty.$$

These indices are defined in relation to the Poisson distribution. The zero-inflated index is easily interpreted, since ZI < 0 indicates zero deflation, ZI $= 0$ corresponds to no excess of zeroes and ZI > 0 indicates ZI. Similarly, HT $\to 1$ when $y \to \infty$ indicates an HT distribution (for a Poisson distribution HT $\to 0$ when $y \to \infty$). Figure 2 presents the dispersion and ZI indices as a function of the expected values μ for different values of the dispersion and Tweedie power parameters. The expected values are defined by $\mu_i = \exp\{\log(10) + 0.8x_i\}$, where x_i is a sequence of length 100 from -1 to 1. We also present the HT index for some extreme values of the random variable. The dispersion parameter was fixed in order to have DI $= 2, 5, 10$ and 20 when the mean equals 10. We refer to these different cases as simulation scenarios 1–4, respectively.

The indices presented in Figure 2 show that for small values of the power parameter, the Poisson–Tweedie distribution is suitable to deal with zero-inflated count data. In that case, the DI and ZI are almost not dependent on the values of the mean. However, the HT decreases as the mean increases. On the other hand, for large values of the power parameter, the HT increases with increasing mean, showing that the model is especially suitable to deal with heavy-tailed count data. In this case, the DI and ZI increase quickly as the mean increases giving an extremely overdispersed model for large values of the mean. In general, the DI and ZI are larger than 1 and 0, respectively, which, of course, shows that the corresponding Poisson–Tweedie distributions cannot deal with underdispersed and zero-deflated count data.

In spite of the integral (2.1) having no simple form in general requiring recursive algorithms for its computation, the first two moments (mean and variance) of the Poisson–Tweedie family can easily be obtained. Jørgensen and Kokonendji (2016) showed by using factorial cumulant function that for $Y \sim PTw_p(\mu, \phi)$, $E(Y) = \mu$ and $\text{Var}(Y) = \mu + \phi\mu^p$. This fact motivates us to specify a model by using only second-order moment assumptions. Thus, consider a cross-section dataset, $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where y_i's are independent and identically distributed (i.i.d.) realizations of Y_i according to an unspecified distribution, whose expectation and variance are given by

$$\begin{aligned} E(Y_i) &= \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \\ \text{Var}(Y_i) &= \text{C}_i = \mu_i + \phi\mu_i^p, \end{aligned} \tag{2.2}$$

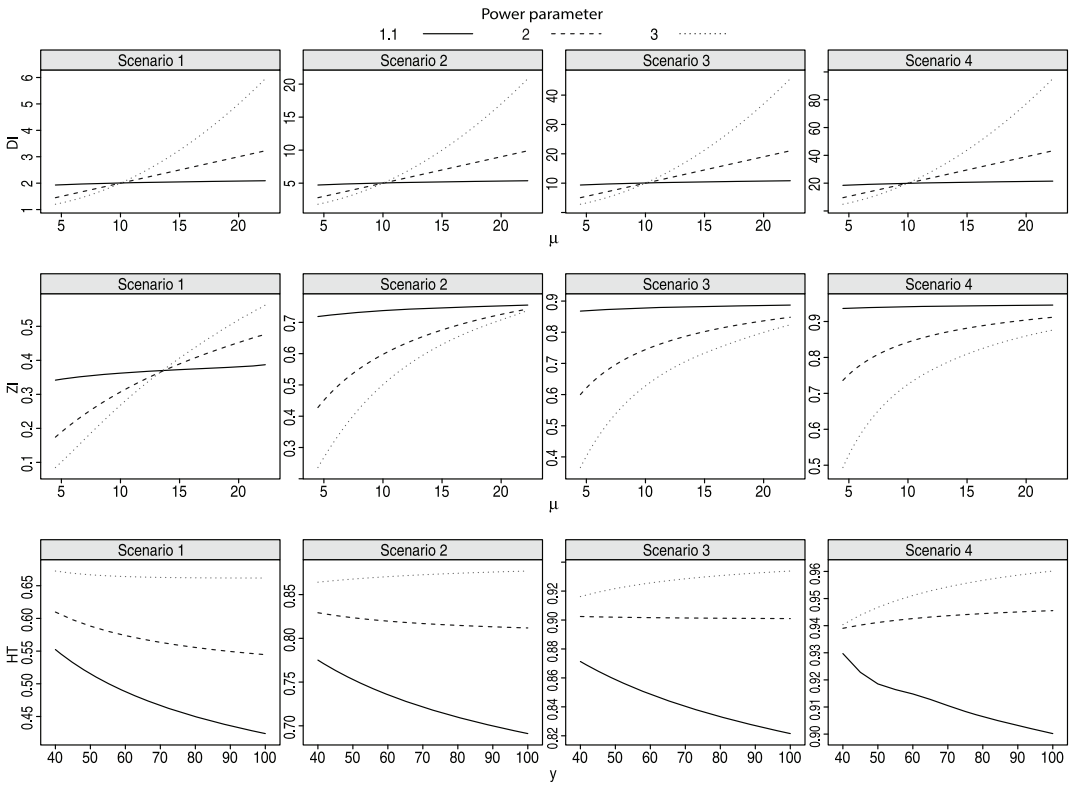

Figure 2 Dispersion (DI) and zero-inflation (ZI) indices as a function of μ by simulation scenarios and Tweedie power parameter values. Heavy tail (HT) index for some extreme values of the random variable Y by simulation scenarios and Tweedie power parameter values

where $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are $(q \times 1)$ vectors of known covariates and unknown regression parameters, respectively. Moreover, g is a standard link function, for which here we adopt the logarithm link function, but potentially any other suitable link function could be adopted. The regression model specified in (2.2) is parametrized by $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\lambda} = (\phi, p)$. Note that, based on second-order moment assumptions, the only restriction to have a proper model is that $\mathrm{Var}(Y_i) > 0$; thus

$$\phi > -\mu_i^{(1-p)},$$

which shows that at least at some extent negative values for the dispersion parameter are allowed. Consequently, the Poisson–Tweedie model can be extended to deal with underdispersed count data; however, in doing so, the associated probability mass functions do not exist. However, in a regression modelling framework, as discussed in this article, we are in general interested in the regression coefficient effects; thus, such an issue does not imply any loss of interpretation and applicability. The formulation

of the extended Poisson–Tweedie model is exactly the same of the quasi-binomial and quasi-Poisson models popular in the context of GLMs; see Nelder and Wedderburn (1972), and McCullagh and Nelder (1989) for details.

Figure 3 presents the DI as a function of the mean for different values of the Tweedie power parameter and negative values for the dispersion parameter. As expected, for negative values of the dispersion parameter, the DI gives values smaller than 1, indicating underdispersion. We also note that, as the mean increases, the DI decreases slowly for small values of the Tweedie power parameter and faster for larger values of the Tweedie power parameter. This shows that the range of negative values allowed for the dispersion parameter decreases rapidly as the value of the Tweedie power parameter increases. Thus, for underdispersed data, we expect small values for the Tweedie power parameter. Furthermore, the second-moment assumptions also allow us to eliminate the non-trivial restriction on the parameter space of the Tweedie power parameter. This makes it possible to estimate values between 0 and 1, where the corresponding Tweedie distribution does not exist. Table 1 presents the main special cases and the dominant features of the extended Poisson–Tweedie models according to the values of the dispersion and power parameters.

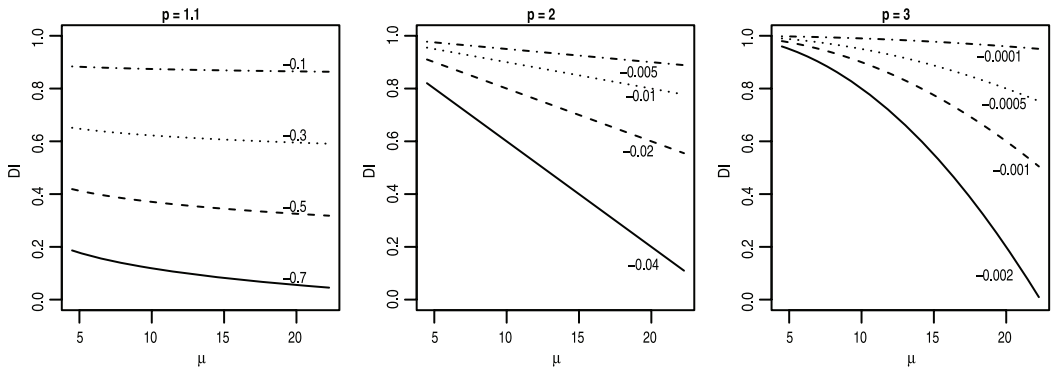

Figure 3 Dispersion index as a function of μ by dispersion and Tweedie power parameter values

Table 1 Reference models and dominant features by dispersion and power parameter values

Reference model	Dominant features	Dispersion	Power
Poisson	Equi	$\phi = 0$	–
Hermite	Over, under	$\phi \lessgtr 0$	$p = 0$
Neyman Type A	Over, under, zero-inflation	$\phi \lessgtr 0$	$p = 1$
Poisson compound Poisson	Over, under, zero-inflation	$\phi \lessgtr 0$	$1 < p < 2$
Pólya–Aeppli	Over, under, zero-inflation	$\phi \lessgtr 0$	$p = 1.5$
Negative binomial	Over, under	$\phi \lessgtr 0$	$p = 2$
Poisson positive stable	Over, heavy tail	$\phi > 0$	$p > 2$
Poisson-inverse Gaussian	Over, heavy tail	$\phi > 0$	$p = 3$

3 Estimation and inference

We shall now introduce the estimating function approach using terminology and results from Jørgensen and Knudsen (2004) and Bonat and Jørgensen (2016). The estimating function approach adopted in this article combines the quasi-score and Pearson estimating functions for the estimation of regression and dispersion parameters, respectively. The quasi-score function for $\boldsymbol{\beta}$ has the following form:

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_1} C_i^{-1} (Y_i - \mu_i), \ldots, \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_q} C_i^{-1} (Y_i - \mu_i) \right)^{\top},$$

where $\partial \mu_i / \partial \beta_j = \mu_i x_{ij}$ for $j = 1, \ldots, q$. The entry (j, k) of the $q \times q$ sensitivity matrix for $\psi_{\boldsymbol{\beta}}$ is given by

$$\mathrm{S}_{\boldsymbol{\beta}_{jk}} = \mathrm{E}\left(\frac{\partial}{\partial \beta_k} \psi_{\boldsymbol{\beta}_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right) = -\sum_{i=1}^{n} \mu_i x_{ij} C_i^{-1} x_{ik} \mu_i. \tag{3.1}$$

In a similar way, the entry (j, k) of the $q \times q$ variability matrix for $\psi_{\boldsymbol{\beta}}$ is given by

$$\mathrm{V}_{\boldsymbol{\beta}_{jk}} = \mathrm{Cov}(\psi_{\boldsymbol{\beta}_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}), \psi_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{\lambda})) = \sum_{i=1}^{n} \mu_i x_{ij} C_i^{-1} x_{ik} \mu_i.$$

Following Jørgensen and Knudsen (2004) and Bonat and Jørgensen (2016), the Pearson estimating function for the dispersion parameters has the following form:

$$\psi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \left(-\sum_{i=1}^{n} \frac{\partial C_i^{-1}}{\partial \phi} \left[(Y_i - \mu_i)^2 - C_i \right], -\sum_{i=1}^{n} \frac{\partial C_i^{-1}}{\partial p} \left[(Y_i - \mu_i)^2 - C_i \right] \right)^{\top}.$$

The Pearson estimating functions are unbiased estimating functions for $\boldsymbol{\lambda}$ based on the squared residuals $(Y_i - \mu_i)^2$ with the expected value C_i.

The entry (j, k) of the 2×2 sensitivity matrix for the dispersion parameters is given by

$$\mathrm{S}_{\boldsymbol{\lambda}_{jk}} = \mathrm{E}\left(\frac{\partial}{\partial \lambda_k} \psi_{\boldsymbol{\lambda}_j}(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right) = -\sum_{i=1}^{n} \frac{\partial C_i^{-1}}{\partial \lambda_j} C_i \frac{\partial C_i^{-1}}{\partial \lambda_k} C_i, \tag{3.2}$$

where λ_1 and λ_2 denote either ϕ or p.

Similarly, the cross entries of the sensitivity matrix are given by

$$\mathrm{S}_{\boldsymbol{\beta}_j \boldsymbol{\lambda}_k} = \mathrm{E}\left(\frac{\partial}{\partial \lambda_k} \psi_{\boldsymbol{\beta}_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right) = 0 \tag{3.3}$$

and

$$S_{\lambda_j \beta_k} = E\left(\frac{\partial}{\partial \beta_k}\psi_{\lambda_j}(\boldsymbol{\lambda}, \boldsymbol{\beta})\right) = -\sum_{i=1}^{n} \frac{\partial C_i^{-1}}{\partial \lambda_j} C_i \frac{\partial C_i^{-1}}{\partial \beta_k} C_i. \tag{3.4}$$

Finally, the joint sensitivity matrix for the parameter vector $\boldsymbol{\theta}$ is given by

$$S_{\boldsymbol{\theta}} = \begin{pmatrix} S_{\boldsymbol{\beta}} & 0 \\ S_{\boldsymbol{\lambda\beta}} & S_{\boldsymbol{\lambda}} \end{pmatrix},$$

whose entries are defined by equations (3.1)–(3.4).

We now calculate the asymptotic variance of the estimating function estimators denoted by $\hat{\boldsymbol{\theta}}$, as obtained from the inverse Godambe information matrix, whose general form for a vector of parameter $\boldsymbol{\theta}$ is $J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-\top}$, where $-\top$ denotes inverse transpose. The variability matrix for $\boldsymbol{\theta}$ has the form

$$V_{\boldsymbol{\theta}} = \begin{pmatrix} V_{\boldsymbol{\beta}} & V_{\boldsymbol{\beta\lambda}} \\ V_{\boldsymbol{\lambda\beta}} & V_{\boldsymbol{\lambda}} \end{pmatrix}, \tag{3.5}$$

where $V_{\boldsymbol{\lambda\beta}} = V_{\boldsymbol{\beta\lambda}}^{\top}$ and $V_{\boldsymbol{\lambda}}$ depend on the third and fourth moments of Y_i, respectively. In order to avoid this dependence on higher order moments, we propose to use the empirical versions of $V_{\boldsymbol{\lambda}}$ and $V_{\boldsymbol{\lambda\beta}}$ as given by

$$\tilde{V}_{\lambda_{jk}} = \sum_{i=1}^{n} \psi_{\lambda_j}(\boldsymbol{\lambda}, \boldsymbol{\beta})_i \psi_{\lambda_k}(\boldsymbol{\lambda}, \boldsymbol{\beta})_i \quad \text{and} \quad \tilde{V}_{\lambda_j \beta_k} = \sum_{i=1}^{n} \psi_{\lambda_j}(\boldsymbol{\lambda}, \boldsymbol{\beta})_i \psi_{\beta_k}(\boldsymbol{\lambda}, \boldsymbol{\beta})_i.$$

Finally, the well-known asymptotic distribution of $\hat{\boldsymbol{\theta}}$ (Jørgensen and Knudsen, 2004; Yuan and Jennrich, 1998; Godambe and Thompson, 1978) is given by

$$\begin{aligned} &\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, J_{\boldsymbol{\theta}}^{-1}), \\ &\text{where} \quad J_{\boldsymbol{\theta}}^{-1} = S_{\boldsymbol{\theta}}^{-1} V_{\boldsymbol{\theta}} S_{\boldsymbol{\theta}}^{-\top}. \end{aligned}$$

To solve the system of equations $\psi_{\boldsymbol{\beta}} = 0$ and $\psi_{\boldsymbol{\lambda}} = 0$, Jørgensen and Knudsen (2004) proposed the modified chaser algorithm, defined by

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - S_{\boldsymbol{\beta}}^{-1} \psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - \alpha S_{\boldsymbol{\lambda}}^{-1} \psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}). \end{aligned}$$

The modified chaser algorithm uses the insensitivity property (3.3), which allows us to use two separate equations to update $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. We introduce the tuning constant, α, to control the step length. A similar version of this algorithm was used by Bonat and Kokonendji (2017) for estimation and inference in the context of Tweedie regression

models. Furthermore, this algorithm is a special case of the flexible algorithm presented by Bonat and Jørgensen (2016) in the context of multivariate covariance GLMs. Hence, estimation for the Poisson–Tweedie model is easily implemented in `R` through the `mcglm` (Bonat, 2016) package.

4 Simulation studies

In this section, we present two simulation studies designed to explore the flexibility of the extended Poisson–Tweedie models to deal with over- and underdispersed count data.

4.1 Fitting extended Poisson–Tweedie models to overdispersed data

In this first simulation study, we designed 12 simulation scenarios to explore the flexibility of the extended Poisson–Tweedie model to deal with overdispersed count data. For each setting, we considered four different sample sizes, 100, 250, 500 and 1 000, generating 1 000 datasets in each case. We considered three values of the Tweedie power parameter, 1.1, 2 and 3, combined with four different degrees of dispersion as measured by the dispersion index. In the case of $p = 1.1$, the dispersion parameter was fixed at $\phi = 0.8, 3.2, 7.2$ and 15. Similarly, for $p = 2$ and $p = 3$, the dispersion parameter was fixed at $\phi = 0.1, 0.4, 0.9, 1.9$ and $\phi = 0.01, 0.04, 0.09, 1.9$, respectively. These values were chosen so that when the mean is 10, the dispersion index takes values of 2, 5, 10 and 20, respectively. The probability mass function of the Poisson–Tweedie distribution for each parameter combination is as presented in Figure 1.

In order to have a regression model structure, we specified the mean vector as $\boldsymbol{\mu}_i = \exp\{\log(10) + 0.8x_{1i} - 1x_{2i}\}$, where x_{1i} is a sequence from -1 to 1 with length equals to the sample size. Similarly, the covariate x_{2i} is a categorical covariate with two levels (0 and 1) and length equals the sample size. Figure 4 shows the average bias plus and minus the average SE for the parameters under each scenario. The scales are standardized for each parameter by dividing the average bias and the limits of the confidence intervals by the SE obtained for the sample of size 100.

The results in Figure 4 show that for all simulation scenarios both the average bias and standard errors tend to 0 as the sample size is increased. This shows the consistency and unbiasedness of the estimating function estimators. Figure 5 presents the confidence interval coverage rate by sample size and simulation scenarios.

The results presented in Figure 5 show that, for the regression parameters, the empirical coverage rates are close to the nominal level of 95% for all sample sizes and simulation scenarios. For the dispersion parameter and a small sample size, the empirical coverage rates are slightly lower than the nominal level, however, they become closer for large samples. On the other hand, for the power parameter, the empirical coverage rates were slightly larger than the nominal level, for all sample sizes and simulation scenarios.

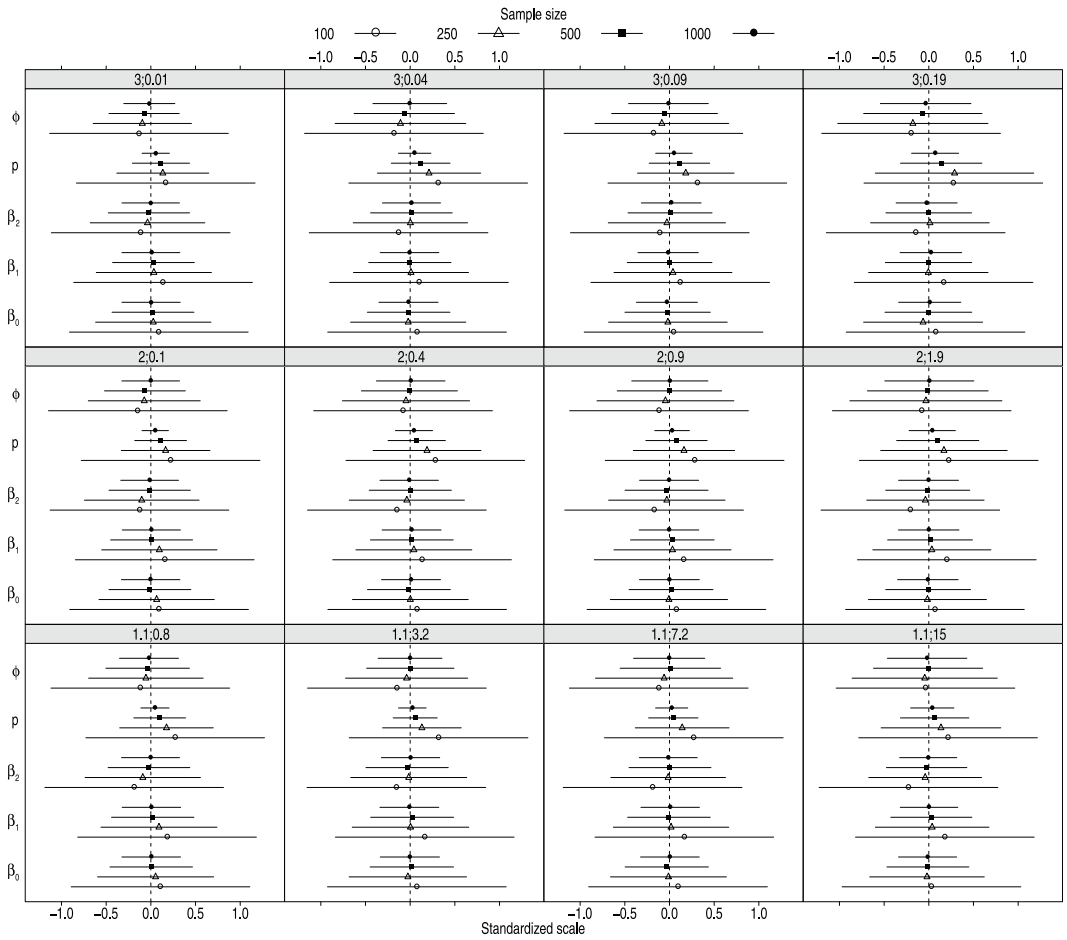

Figure 4 Average bias and confidence intervals on a standardized scale by sample size and simulation scenario

4.2 Fitting extended Poisson–Tweedie models to underdispersed data

As discussed in Section 2, the extended Poisson–Tweedie model can deal with underdispersed count data by allowing negative values for the dispersion parameter. However, in that case, there is no probability mass function associated with the model. Consequently, it is impossible to use such a model to simulate underdispersed data. Thus, we simulated datasets from the COM-Poisson (Sellers and Shmueli, 2010) and Gamma-count (Zeviani et al., 2014) distributions. Such models are well known in the literature for their ability to model underdispersed data.

Following the parametrization used by Sellers and Shmueli (2010), $Y \sim CP(\lambda, \nu)$ denotes a COM-Poisson distributed random variable. Similarly, we write $Y \sim GC(\lambda, \nu)$ for a Gamma-count distributed random variable. For both distributions,

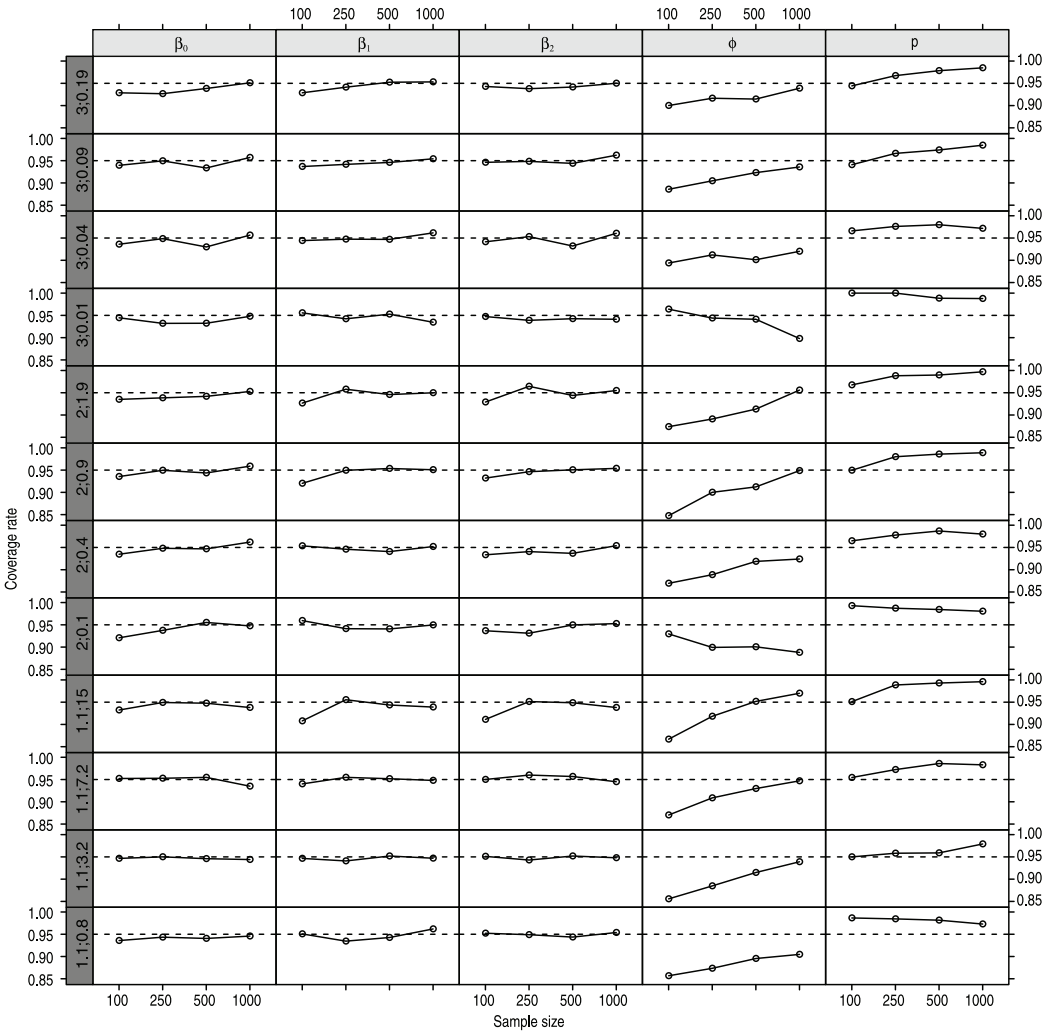

Figure 5 Coverage rate for each parameter by sample size and simulation scenarios

the additional parameter ν controls the dispersion structure, with values larger than 1 indicating underdispersed count data. An inconvenience of the COM-Poisson and Gamma-count regression models as proposed by Sellers and Shmueli (2010) and Zeviani et al. (2014), respectively, is that the regression structure is not linked to a function of $E(Y)$, as is usual in the GLM framework. To overcome this limitation and obtain parameters that are interpretable in the usual way, that is, related directly to a function of $E(Y)$, we take an alternative approach based on simulation. The procedure consisted of specifying the λ parameter using a regression structure, $\lambda_i = \exp\{\lambda_0 + \lambda_1 x_1\}$ for $i = 1, \ldots, n$, where n denotes the sample size and x_1 is a sequence from -1 to 1 and length n. For each value of λ, we simulate 1 000 values

and compute the empirical mean and variance. We denote these quantities by $\widehat{E(Y)}$ and $\widehat{\text{var}(Y)}$. Then, we fitted two nonlinear models specified as $\widehat{E(Y)} = \exp(\beta_0 + \beta_1 x_1)$ and $\widehat{\text{var}(Y)} = \widehat{E(Y)} + \phi\widehat{E(Y)}^p$. From these fits, we obtained the expected values of the regression, dispersion and Tweedie power parameters.

We designed four simulation scenarios by introducing different degrees of underdispersion in the datasets. The parameter ν was fixed at the values $\nu = 2, 4, 6$ and 8 for both distributions. In the COM-Poisson case, we took $\lambda_0 = 8$ and $\lambda_1 = 4$, and for the Gamma-count case, we fixed $\lambda_0 = 2$ and $\lambda_1 = 1$. It is important to highlight that, for all of these selected values, the expected value of the dispersion parameter ϕ is negative. The particular values depend on λ_0, λ_1 and ν and are presented for both distributions in Table 2.

Table 2 Corresponding values of β_0, β_1, ϕ and p depending on the values of λ_0, λ_1 and ν for the COM-Poisson and Gamma-count distributions

			COM-Poisson			
ν	λ_0	λ_1	β_0	β_1	ϕ	p
2	8	4	3.995	2.004	−0.485	1.008
4	8	4	1.941	1.047	−0.714	1.014
6	8	4	1.206	0.744	−0.790	1.020
8	8	4	0.803	0.602	−0.821	1.036
			Gamma-count			
2	2	1	1.962	1.028	−0.429	1.045
4	2	1	1.943	1.042	−0.682	1.003
6	2	1	1.936	1.048	−0.779	1.019
8	2	1	1.932	1.051	−0.820	1.020

For each setting, we generated 1 000 datasets for four different sample sizes 100, 250, 500 and 1 000. The extended Poisson–Tweedie model was fitted using the estimating function approach presented in the Section 3. Figure 6 shows the average bias plus and minus the average SE for the parameters in each scenario. For each parameter, the scales are standardized by dividing the average bias and limits of the confidence intervals by the SE obtained for the sample of size 100.

The results in Figure 4 show that for all simulation scenarios, both the average bias and SEs tend to 0 as the sample size is increased for both dispersion and Tweedie power parameters. It shows the consistency of the estimating function estimators. Concerning the regression parameters, in general, the intercept (β_0) is underestimated, while the slope (β_1) is overestimated. The bias is larger for the Gamma-count data with strong underdispersion ($\nu = 8$) case. However, it is still small in its magnitude.

5 Data analyses

In this section, we present four examples to illustrate the application of the extended Poisson–Tweedie models. The data and the `R` scripts used for their analysis

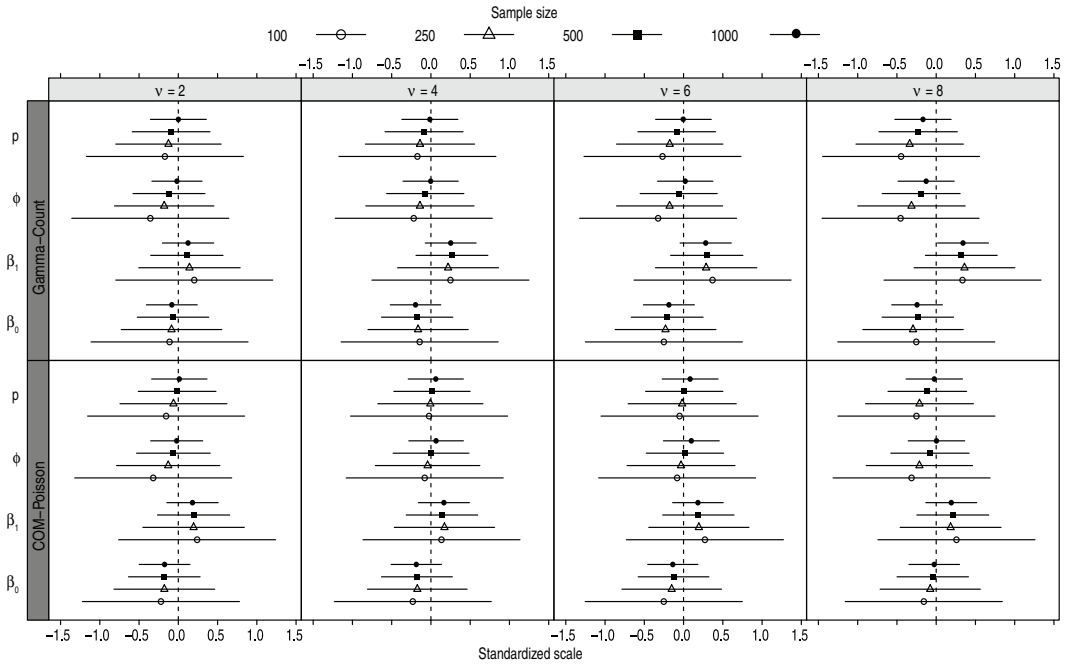

Figure 6 Average bias and confidence interval on a standardized scale by sample size and simulation scenario

can be obtained from `http://www.leg.ufpr.br/doku.php/publications:papercompanions:ptw`.

5.1 Dataset 1: Respiratory disease morbidity among children in Curitiba, Paraná, Brazil

The first example concerns monthly morbidity from respiratory diseases among 0–4-year-old children in Curitiba, Paraná State, Brazil. The data were collected for the period from January 1995 to December 2005, corresponding to 132 months. The main goal of the investigation was to assess the effect of three environmental covariates (precipitation, maximum and minimum temperatures) on the morbidity from respiratory diseases. Figure 7 presents a time series plot with fitted values (A) and dispersion diagrams of the monthly morbidity from respiratory diseases against the covariates precipitation (B), maximum temperature (C) and minimum temperature (D), with a simple linear fit indicated by the straight black lines. These plots indicate a clear seasonal pattern and the essentially linear effect of all covariates (as suggested by the simple linear fits superimposed in Figure 7). The linear predictor is expressed in terms of Fourier harmonics (seasonal variation) and the effect of the three environmental covariates. The logarithm of the population size was used as an offset. To compare the extended Poisson–Tweedie model with the

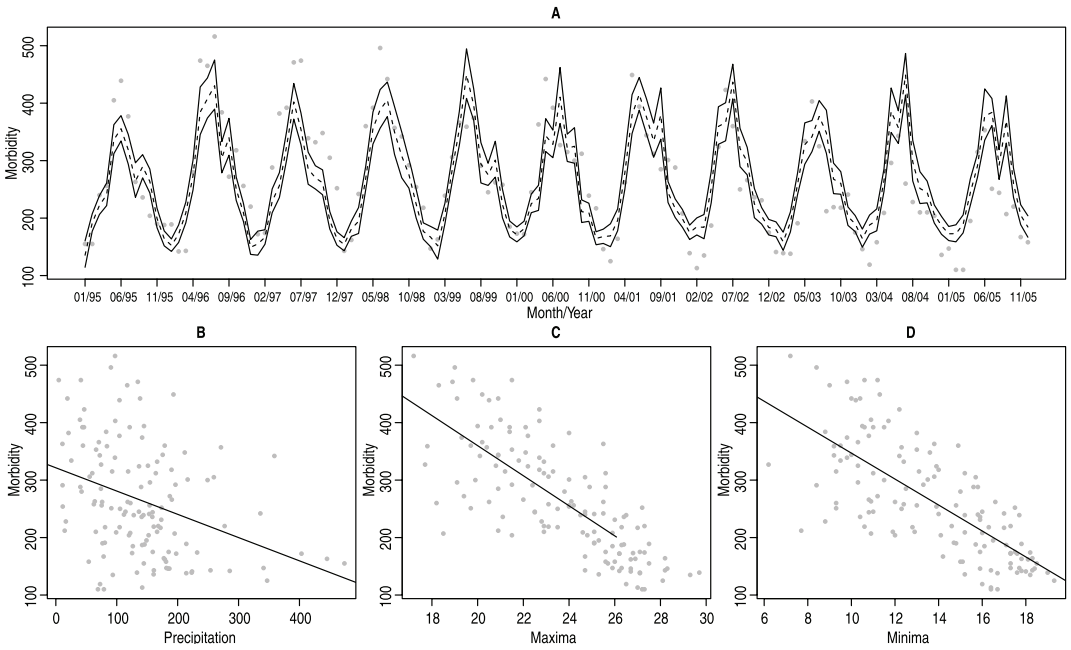

Figure 7 Time series plot with fitted values (A) and dispersion diagrams of the monthly morbidity by respiratory diseases against the covariates precipitation (B), maximum temperature (C) and minimum temperature (D), with a simple linear fit indicated by the straight black lines

usual Poisson log-linear model, Table 3 shows the corresponding estimates and SEs, along with the ratios between the both model estimates and SEs.

The results presented in Table 3 show that the estimates from the extended Poisson–Tweedie and Poisson models are similar. However, the SE from the extended Poisson–Tweedie model are, in general, 3.5 times larger than the ones from the Poisson model. This difference is explained by the dispersion structure. The dispersion parameter $\phi > 0$ indicates overdispersion, which implies that the SEs obtained by the Poisson model are underestimated. The Poisson model gives evidence of a significant effect for all covariates, while the Poisson–Tweedie model only gives significant effects for the seasonal variation and the temperature maxima covariates. The fitted values and 95% confidence interval are shown in Figure 7(A). The model captures the swing in the data and highlights the seasonal behaviour with high and low morbidity numbers around winter and summer months, respectively. The negative effect of the covariate temperature maxima agrees with the seasonal effects and the exploratory analysis presented in Figure 7(C). The power parameter estimate with its corresponding SE indicates that all Poisson–Tweedie models with $p \in [1, 2]$ are suitable for this dataset. In particular, NTA, Pólya–Aeppli and negative binomial distributions can be good choices.

Table 3 Dataset 1: Parameter estimates and standard errors (SEs) for Poisson–Tweedie and Poisson models (first and second columns). Ratios between Poisson–Tweedie and Poisson estimates and SEs (third column)

	Estimates (SE)		
Parameter	Poisson–Tweedie	Poisson	Ratio
`Intercept`	2.277 (0.304)*	2.226 (0.084)*	1.023 (3.598)
`cos(2*pi*Month/12)`	−0.223 (0.056)*	−0.226 (0.016)*	0.985 (3.507)
`sin(2*pi*Month/12)`	−0.093 (0.048)*	−0.073 (0.013)*	1.279 (3.562)
`Maxima`	−0.083 (0.017)*	−0.083 (0.005)*	1.057 (3.590)
`Minima`	0.039 (0.022)	0.034 (0.006)*	1.128 (3.592)
`Precipitation`	−0.001 (0.000)	−0.001 (0.000)*	0.978 (3.337)
p	1.652 (0.423)	–	–
ϕ	0.293 (0.036)	–	–

5.2 Dataset 2: Cotton bolls greenhouse experiment

The second example relates to cotton boll production and is from a completely randomized experiment conducted in a greenhouse. The aim was to assess the effect of five artificial defoliation levels (0%, 25%, 50%, 75% and 100%) and five growth stages (vegetative, flower bud, blossom, fig and cotton boll) on the number of cotton bolls. There were five replicates of each treatment combination, giving a dataset with 125 observations. This dataset was analysed in Zeviani et al. (2014) using the Gamma-count distribution, since there was clear evidence of underdispersion. Following Zeviani et al. (2014), the linear predictor was specified by

$$g(\mu_{ij}) = \beta_0 + \beta_{1j}\mathbf{def}_i + \beta_{2j}\mathbf{def}_i^2,$$

where μ_{ij} is the expected number of cotton bolls for the defoliation (**def**) level $i = 1, \ldots, 5$ and growth stage $j = 1, \ldots, 5$, that is, we have a second-order effect of defoliation in each growth stage. Table 4 presents the estimates and SEs for the Poisson–Tweedie and standard Poisson models, along with the ratios between the respective estimates and SEs.

The results in Table 4 show that the estimates are quite similar; however, the SE obtained by the Poisson–Tweedie model are smaller than those from the Poisson model. This is explained by the negative estimate of the dispersion parameter, which indicates underdispersion. The value of the power parameter is close to 1 and explains the similarity of the regression parameter estimates. Appropriate estimation of the SE is important for this dataset, since the Poisson–Tweedie identifies the effect of the defoliation as significant for three of the five growth stages, while the Poisson model only finds the defoliation effect as significant for the blossom growth stage. Figure 8 presents the observed values and curves of fitted values (Poisson in grey and Poisson–Tweedie in black) and confidence intervals (95%) as functions of the defoliation level for each growth stage and supports the preceding conclusions.

The results from the Poisson–Tweedie model are consistent with those from the Gamma-count model, fitted by Zeviani et al. (2014), in that both methods indicate underdispersion and significant effects of defoliation for the vegetative, blossom and fig growth stages. However, it is important to note that the estimates obtained by the Gamma-count model fitted by Zeviani et al. (2014) are not directly comparable with the ones obtained from the Poisson–Tweedie model, since the latter is modelling the expectation, while the Gamma-count distribution models the distribution of the time between events.

Table 4 Dataset 2: Parameter estimates and SEs for Poisson–Tweedie and Poisson models (first and second columns). Ratios between Poisson–Tweedie and Poisson estimates and SEs (third column)

	Estimates (SE)		
Parameter	Poisson–Tweedie	Poisson	Ratio
`Intercept`	2.189 (0.030)*	2.190 (0.063)*	1.000 (0.471)
`vegetative:des`	0.438 (0.243)	0.437 (0.516)	1.003 (0.471)
`vegetative:des`^{2}	−0.806 (0.274)*	−0.805 (0.584)	1.001 (0.469)
`flower bud:des`	0.292 (0.239)	0.290 (0.508)	1.007 (0.471)
`flower bud:des`^{2}	−0.490 (0.266)	−0.488 (0.566)	1.004 (0.470)
`blossom:des`	−1.235 (0.281)*	−1.242 (0.604)*	0.994 (0.465)
`blossom:des`^{2}	0.665 (0.316)*	0.673 (0.680)	0.989 (0.465)
`fig:des`	0.380 (0.265)	0.365 (0.566)	1.040 (0.468)
`fig:des`^{2}	−1.330 (0.313)*	−1.310 (0.673)	1.015 (0.465)
`boll:des`	0.011 (0.237)	0.009 (0.504)	1.181 (0.471)
`boll:des`^{2}	−0.021 (0.260)	−0.020 (0.553)	1.059 (0.471)
p	0.981 (0.137)	–	–
ϕ	−0.810 (0.223)	–	–

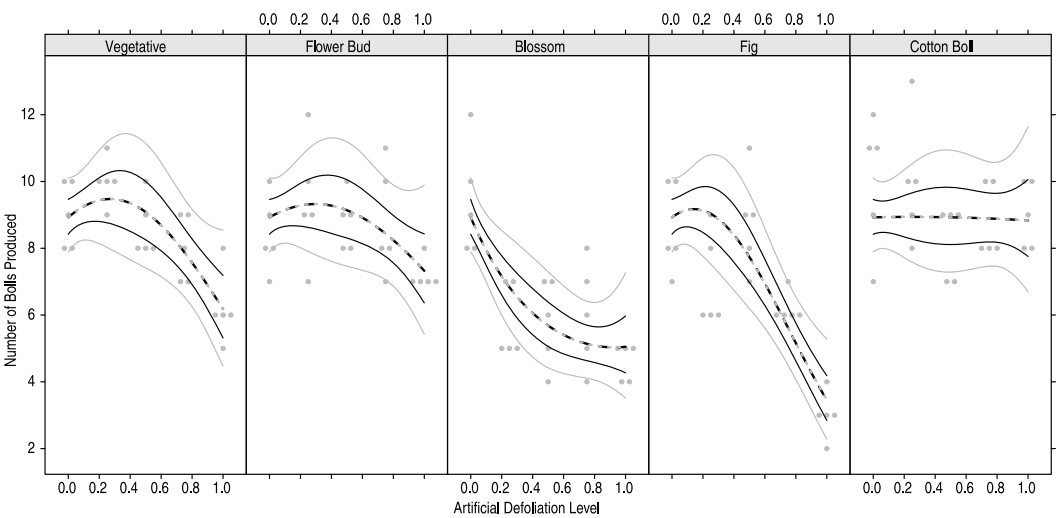

Figure 8 Dispersion diagrams of observed values and curves of fitted values (Poisson-grey and Poisson–Tweedie-black) and confidence intervals (95%) as functions of the defoliation level for each growth stage

5.3 Dataset 3: Radiation-induced chromosome aberration counts

In this example, we apply the extended Poisson–Tweedie model to describe the number of chromosome aberrations in biological dosimetry. The dataset considered was obtained after irradiating blood samples with five different doses between 0.1 and 1 Gy of 2.1 MeV neutrons. In this case, the frequencies of dicentrics and centric rings after a culture of 72 hours are analysed. The dataset in Table 5 was first presented by Heimers et al. (2006) and analysed by Oliveira et al. (2016) as an example of zero-inflated data.

We fitted the extended Poisson–Tweedie and Poisson models with the linear predictor specified as a quadratic dose model, that is,

$$g(\mu_{ij}) = \beta_0 + \beta_1 \mathtt{dose}_i + \beta_2 \mathtt{dose}_i^2.$$

Table 6 presents the estimates and SEs for the Poisson–Tweedie and Poisson models, along with the ratios between the respective estimates and SEs.

Results in Table 6 show evidence of weak overdispersion that can be attributed to ZI, since the estimate of the power parameter was close to 1, which in turn implies that the SE obtained from the Poisson–Tweedie model are around 10% larger than those obtained from the Poisson model.

For this dataset, it is particularly easy to compute the log-likelihood value, since we have only a few unique observed counts and dose values. Thus, we can use log-likelihood values to compare the fit of the Poisson–Tweedie model with the fit obtained by the zero-inflated Poisson and zero-inflated negative binomial models. The log-likelihood value of the Poisson–Tweedie model was −2 950.605, while the

Table 5 Frequency distributions of the number of dicentrics and centric rings by dose levels

	y_{ij}							
x_i	0	1	2	3	4	5	6	7
0.1	2 281	130	21	1	0	0	0	0
0.3	847	127	19	6	1	0	0	0
0.5	567	165	49	16	2	0	0	0
0.7	356	167	62	9	5	1	0	0
1	169	131	72	18	9	0	0	1

Table 6 Dataset 3: Parameter estimates and SEs for Poisson–Tweedie and Poisson models (first and second columns). Ratios between Poisson–Tweedie and Poisson estimates and SEs (third column)

	Estimates (SE)		
Parameter	Poisson–Tweedie	Poisson	Ratio
`Intercept`	−3.126 (0.106)*	−3.125 (0.097)*	1.000 (1.098)
`dose`	5.514 (0.408)*	5.508 (0.369)*	1.001 (1.104)
$\mathtt{dose}^2$	−2.481 (0.342)*	−2.476 (0.309)*	1.002 (1.107)
p	1.085 (0.299)	–	–
ϕ	0.249 (0.100)	–	–

Table 7 Dataset 4: Estimates and SEs from different models

Parameter	Poisson	NTA	NB	PIG
`Intercept`	2.942 (0.207)*	2.942 (0.194)*	2.937 (0.197)*	2.933 (0.203)*
`nhu`	0.061 (0.014)*	0.061 (0.013)*	0.060 (0.013)*	0.060 (0.014)*
`aid`	−0.012 (0.002)*	−0.012 (0.002)*	−0.012 (0.002)*	−0.012 (0.002)*
`aha`	−0.004 (0.002)*	−0.004 (0.002)*	−0.004 (0.002)*	−0.004 (0.002)*
`dnc`	0.168 (0.026)*	0.168 (0.024)*	0.165 (0.025)*	0.166 (0.025)*
`ds`	−0.129 (0.016)*	−0.129 (0.015)*	−0.127 (0.015)*	−0.127 (0.016)*
ϕ	0	−0.122(0.123)	−0.008 (0.010)	0.000 (0.000)
p	–	1	2	3

maximized log-likelihood value of the zero-inflated Poisson and zero-inflated negative binomial models were −2 950.462 and −2 950.531, respectively. Furthermore, the maximized log-likelihood value of the Poisson model was −2 995.389. These results show that the Poisson–Tweedie model can offer a very competitive fit, even without an additional linear predictor to describe the excess of zeroes. Furthermore, it is interesting to note that in spite of the large difference in the log-likelihood values, the Poisson model provides the same interpretation in terms of the significance of the covariates as the Poisson–Tweedie model for this dataset.

5.4 Dataset 4: Customers' profile

The last example corresponds to a dataset collected to investigate the customer profile of a large company of household supplies. During a representative two-week period, in-store surveys were conducted and addresses of customers were obtained. The addresses were then used to identify the metropolitan area census tracts in which the customers resident. At the end of the survey period, the total number of customers who visited the store from each census tract within a 10-mile radius was determined and relevant demographic information for each tract was obtained. The dataset was analysed in Neter et al. (1996) as an example of Poisson regression model, since it is a classic example of equidispersed count data. Following Neter et al. (1996), we considered the covariates, number of housing units (**`nhu`**), average income in dollars (**`aid`**), average housing unit age in years (**`aha`**), distance to the nearest competitor in miles (**`dnc`**) and distance to store in miles (**`ds`**) for forming the linear predictor.

For equidispersed data, the estimation of the Tweedie power parameter is in general a difficult task. In this case, the dispersion parameter ϕ should be estimated around zero. Thus, we do not have enough information to distinguish between different values of the Tweedie power parameter. Consequently, we can fix the Tweedie power parameter at any value, and the corresponding fitted models should be very similar. To illustrate this idea, we fitted the extended Poisson–Tweedie model fixing the Tweedie power parameter at the values 1, 2 and 3, corresponding to the NTA, NB and PIG distributions, respectively. We also fitted the standard Poisson model for comparison, the estimates and SEs are presented in Table 7.

The results presented in Table 7 clearly show that for all fitted models, the dispersion parameter does not differ from zero, which gives evidence of equidispersion. The regression coefficients and the associated SE do not depend on the models, and in particular do not depend on the power parameter value. This example shows that, although a more careful analysis is required, the extended Poisson–Tweedie model can deal with equidispersed data. Furthermore, the estimation of the extra dispersion parameter does not inflate the SE associated with the regression coefficients. Thus, there is no loss of efficiency when using the Poisson–Tweedie model for equidispersed count data.

6 Discussion

We presented a flexible statistical modelling framework to deal with count data. The models are based on the Poisson–Tweedie family of distributions that automatically adapts to overdispersed, zero-inflated and heavy-tailed count data. Furthermore, we adopted an estimating function approach for estimation and inference based only on second-moment assumptions. Such a specification allows us to extend the Poisson–Tweedie model to deal with underdispersed count data by allowing negative values for the dispersion parameter. The main technical advantage of the second-order moment specification is the simplicity of the fitting algorithm, which amounts to finding the root of a set of nonlinear equations. The Poisson–Tweedie family encompasses some of the most popular models for count data, such as the Hermite, NTA, Pólya–Aeppli, negative binomial and PIG distributions. For this reason, the estimation of the power parameter plays an important role in the context of Poisson–Tweedie regression models, since it is an index that distinguishes between these important distributions. Thus, the estimation of the power parameter can work as an automatic distribution selection.

The modified chaser algorithm depends upon good initial values for fast and reliable convergence. To obtain initial values for the regression coefficients, we recommend using parameter estimates from the fit of a standard Poisson regression model. For the dispersion parameter, we recommend using the Pearson estimator, and for the initial value for the power parameter, we suggest taking a value of 1.

Regarding the convergence of the algorithm, there are two cases where the convergence is particularly challenging. First, in the case of underdispersion, we expect negative values for the dispersion parameter, and in this case, the parameter space is not trivially defined. Consequently, the estimate can be very close to the border of the parameter space, and in such cases, the algorithm can have convergence problems. To handle this situation, we introduced the tuning constant parameter α in the update for λ, which allows us to control the step length and avoid values outside of the parameter space. Second, in the case of equidispersion, we expect $\phi \approx 0$, so that although this is not a value on the border of the parameter space, it makes it impossible to estimate the power parameter p. Thus, in this case, the power parameter should be fixed at some value. This situation was illustrated in our fourth example.

We conducted a simulation study on the properties of the estimating function estimators. The results showed that in general the estimating function estimators are unbiased and consistent. We also evaluated the validity of the SE obtained by the estimating function approach by computing the empirical coverage rate. The results showed that for the regression coefficients, our estimators provide the specified level of coverage for all simulation scenarios and sample sizes. Regarding the dispersion parameter, the results showed that for small samples the SE are underestimated; however, the results improve for larger samples. On the other hand, the SE associated with the power parameter are overestimated for all simulation scenarios and sample sizes. However, the coverage rate presented values only slightly larger than the specified nominal level of 95%. It is important to highlight that the under- or overestimation of the dispersion and power parameters do not affect the estimates and SE associated with the regression coefficients. This is due to the insensitivity property; see equation (3.3). Furthermore, we demonstrated the flexibility of the extended Poisson–Tweedie model to deal with underdispersed count data as generated by the COM-Poisson and Gamma-count distribution. It also shows that the model has a good level of robustness against model misspecification.

Discussion of the efficiency of the estimating function estimators is difficult due to the lack of a closed form for the Fisher information matrix. Bonat and Kokonendji (2017) showed in the context of Tweedie regression models that the quasi-score function provides asymptotically efficient estimators for the regression parameters; thus, a similar result is expected for the Poisson–Tweedie regression model. Concerning the dispersion and power parameters, the fact that the sensitivity and variability matrices do not coincide indicates that the Pearson estimating functions are not optimum. Furthermore, the use of empirical third and fourth moments for the calculation of the Godambe information matrix must imply some efficiency loss. On the other hand, it again makes the model robust against misspecification.

We analysed four real datasets to explore and illustrate the flexibility of the extended Poisson–Tweedie model. Dataset 1 presented a classical case of overdispersion. This dataset illustrated the most common problem when using the Poisson model for overdispersed count data, that is, the strong underestimation of the SE associated with the regression coefficients. The Poisson–Tweedie model automatically adapts to the dispersion in the data by the estimation of the dispersion parameter, while choosing the appropriate distribution in the Poisson–Tweedie family through the estimation of the power parameter. Furthermore, the uncertainty around the data distribution is taken into account and can be assessed based on the SE associated with the power parameter. In particular, for this application, the model shows that any distribution in the family of the Poisson compound Poisson distributions ($1 < p < 2$) provides a suitable fit for the dataset. Thus, we avoid the need to fit an array of models and the use of measures of goodness of fit to choose between them.

Dataset 2 presents the less frequent case of underdispersion. In this case, the problem is that the Poisson model overestimates the SE associated with the regression coefficients. The negative value of the dispersion parameter obtained by fitting the Poisson–Tweedie model to this dataset indicates underdispersion. Thus, the model

automatically corrects the SE for the regression coefficients, giving SEs that are smaller than those obtained from the Poisson model. The problem of zero-inflated count data was illustrated by the dataset 3. In this example, we showed that, in general, ZI introduces overdispersion and that the Poisson–Tweedie model can also adapt to ZI providing a very competitive fit when compared with more orthodox approaches such as the zero-inflated Poisson and zero-inflated negative binomial models. Finally, dataset 4 illustrated the case of equidispersed count data. This case is particularly challenging for the Poisson–Tweedie model since the dispersion parameter should be zero, which implies that any distribution in the family of Poisson–Tweedie distributions can provide a suitable fit for the data. Thus, the estimation of the Tweedie power parameter is very difficult, because the estimating function associated with the Tweedie power parameter is flat. In this case, our approach was to fit the model with the Tweedie power parameter fixed at the values 1, 2 and 3. We compared the fit of these three models with the fit of the Poisson model and, since we have equidispersed data, all models provided quite similar estimates and SEs. Furthermore, all models indicated that the dispersion parameter is not different from zero, which again indicates equidispersion. It is important to emphasize that the estimation of the additional dispersion parameter does not inflate the SEs associated with the regression parameters.

There are many possible extensions to the basic model discussed in the present article, including incorporating penalized splines and the use of regularization for high dimensional data, with important applications in genetics. There is also a need to develop methods for model checking such as residual analysis, leverage and outlier detection. Finally, we can extend the model to deal with multivariate count data, with many potential applications for the analysis of longitudinal and spatial data. These extensions will form the basis of future work.

Acknowledgements

This article is dedicated in honour and memory of Professor Bent Jørgensen. This work was done while the first author was visiting the Laboratory of Mathematics of Besançon, France and School of Mathematics of National University of Ireland, Galway, Ireland. The first author is supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil. The last author was partially supported by CNPq, a Brazilian science funding agency.

Supplementary material

`http://www.leg.ufpr.br/doku.php/publications:papercompanions:ptw`

References

Barabesi L, Becatti C and Marcheselli M (2016) *The tempered discrete Linnik distribution*. ArXiv e-prints 1605.02326.

Bonat WH (2016) *Mcglm: Multivariate covariance generalized linear models*. R package version 0.3.0. URL http://git.leg.ufpr.br/wbonat/mcglm (last accessed 1 August 2017).

Bonat WH and Jørgensen B (2016) Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **65**, 649–75.

Bonat WH and Kokonendji CC (2017) Flexible Tweedie regression models for continuous data. *Journal of Statistical Computation and Simulation* **87**, 2138–52.

Dunn PK (2013) *Tweedie: Tweedie exponential family models*. R package version 2.1.7. URL https://CRAN.R-project.org/package=tweedie (last accessed 1 August 2017).

El-Shaarawi AH, Zhu R and Joe H (2011) Modelling species abundance using the Poisson–Tweedie family. *Environmetrics* **22**, 152–64.

Esnaola M, Puig P, Gonzalez D, Castelo R and Gonzalez JR (2013) A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated rna-seq experiments. *BMC Bioinformatics* **14**, 254–76.

Godambe VP and Thompson M (1978) Some aspects of the theory of estimating equations. *Journal of Statistical Planning and Inference* **2**, 95–104.

Heimers A, Brede HJ, Giesen U and Homann W (2006) Chromosome aberration analysis and the influence of mitotic delay after simulated partial-body exposure with high doses of sparsely and densely ionising radiation. *Radiation and Environmental Biophysics* **45**, 45–54.

Hinde J and Demétrio CGB (1998) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis* **27**, 151–70.

Jørgensen B (1987) Exponential dispersion models. *Journal of the Royal Statistical Society, Series B (Methodological)* **49**, 127–62.

Jørgensen B (1997) *The theory of dispersion models*. London: Chapman and Hall.

Jørgensen B and Knudsen SJ (2004) Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics* **31**, 93–114.

Jørgensen B and Kokonendji CC (2016) Discrete dispersion models and their Tweedie asymptotics. *AStA Advances in Statistical Analysis* **100**, 43–78.

Kalktawi HS, Vinciotti V and Yu K (2015) *A simple and adaptive dispersion regression model for count data*. ArXiv e-prints 1511.00634.

Kokonendji CC, Demétrio, CGB and Zocchi SS. (2007). On Hinde–Demétrio regression models for overdispersed count data. *Statistical Methodology* **4**, 277–91.

Kokonendji CC, Dossou-Gbete S and Demétrio CGB (2004) Some discrete exponential dispersion models: Poisson–Tweedie and Hinde–Demétrio classes. *Statistics and Operations Research Transactions* **28**, 201–14.

Liang KY and Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Loeys T, Moerkerke B, De Smet O and Buysse A (2012) The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology* **65**, 163–80.

McCullagh P and Nelder J (1989) *Generalized linear models, 2nd edition* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). London: Taylor & Francis.

Nelder JA and Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**, 370–84.

Neter J, Kutner MH, Nachtsheim CJ and Wasserman W (1996) *Applied linear statistical models*. Chicago, IL: Irwin.

Oliveira M, Einbeck J, Higueras M, Ainsbury E, Puig P and Rothkamm K (2016) Zero-inflated regression models for radiation-induced chromosome aberration data: A comparative study. *Biometrical Journal* **58**, 259–79.

R Core Team (2016). R Core Team (2016) R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www. R-project.org/ (last accessed 1 August 2017).

Ridout MS, Demétrio CGB and Hinde JP (1998) Models for count data with many zeros. *Proceedings of the XIXth International Biometrics Conference*, Cape Town, South Africa.

Rigby RA, Stasinopoulos DM and Akantziliotou C (2008) A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics and Data Analysis* **53**, 381–93.

Sellers KF and Raim A (2016) A flexible zero-inflated model to address data dispersion. *Computational Statistics and Data Analysis* **99**, 68–80.

Sellers KF and Shmueli G (2010) A flexible regression model for count data. *Annals of Applied Statistics* **4**, 943–61.

Smyth GK and Jørgensen B (2002) Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin: The Journal of the International Actuarial Association* **32**, 143–57.

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika* **61**, 439–47.

Yuan K-H and Jennrich RI (1998) Asymptotics of estimating equations under natural conditions. *Journal of Multivariate Analysis* **65**, 245–60.

Zeger SL, Liang K-Y and Albert PS (1988) Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–60.

Zeileis A, Kleiber C and Jackman S (2008) Regression models for count data in R. *Journal of Statistical Software* **27**, 1–25.

Zeviani WM, Ribeiro Jr PJ, Bonat WH, Shimakura SE and Muniz JA (2014) The gamma-count distribution in the analysis of experimental underdispersed data. *Journal of Applied Statistics* **41**, 2616–26.

Zhu R and Joe H (2009) Modelling heavy-tailed count data using a generalised Poisson-inverse Gaussian family. *Statistics and Probability Letters* **79**, 1695–1703.