Revista Colombiana de Estadística - Applied Statistics
January 2024, volume 47, issue 1, pp. 67–86
http://doi.org/10.15446/rce.v47n1.101952

Poisson-Tweedie Models for Count Data with Excessive Zeros. Comparison with the Negative Binomial Model

Modelos Poisson-Tweedie para datos de conteo con exceso de ceros. Comparación con el modelo binomial negativo

Guillermina B. Harvey[a], Gabriela S. Boggio[b]

Escuela de Estadística, Universidad Nacional de Rosario, Rosario, Argentina

Abstract

The presence of a large number of zero counts is quite common in studies involving count data. This causes overdispersion. Therefore, different types of models have been proposed as alternatives and a very frequent practice is to use the negative binomial model. In 2018, Bonat (2018) considered a new type of model, based on the Poisson-Tweedie dispersion models, which can automatically adapt to different degrees of overdispersion in count data. This article presents a simulation study in order to compare the estimates derived from the Poisson-Tweedie model for a wide range of overdispersed data with estimates derived from the negative binomial model. In both models, the relative percent bias of the estimated coefficients was very small. Nevertheless, the Poisson-Tweedie model showed a better performance with smaller values for the mean squared errors, particularly in scenarios with more dispersion. Hence, it would be possible to suggest the data analyst in which situations it would be enough to work with the popular negative binomial model or when it would be best to use the Poisson-Tweedie family. Additionally, the comparison between the fit of the negative binomial model and that of the Poisson-Tweedie family is illustrated by analysing the number of pediatric consultations of a group of children who receive health care in a public health center in Rosario, Argentina. Although the results obtained in both models were similar, the estimates in the Poisson-Tweedie model were more accurate.

Key words: Count data; Poisson-Tweedie models; Zero-inflation.

[a]Master. E-mail: gharvey@fcecon.unr.edu.ar
[b]Ph.D. E-mail: gboggio@fcecon.unr.edu.ar

Resumen

En estudios que involucran el análisis de datos de conteo es común encontrar una gran cantidad de ceros. La sobredispersión que ello provoca ha sido tenida en cuenta en diferentes alternativas de modelización siendo el modelo binomial negativo la más utilizada. En 2018 se suma la propuesta desarrollada por Bonat (2018) ellos consideraron una nueva clase de modelos, basada en los modelos con dispersión Poisson-Tweedie, los cuales se adaptan en forma automática a diferentes grados de sobredispersión en datos de conteo. Este trabajo presenta un estudio por simulación para comparar las estimaciones derivadas del modelo Poisson-Tweedie con las del binomial negativo frente a diferentes niveles de sobredispersión. Se encontraron estimaciones de los coeficientes del modelo con sesgos muy pequeños para ambos modelos y errores cuadráticos medios levemente menores para el modelo Poisson-Tweedie, evidenciando su mejor desempeño en los escenarios de mayor dispersión. Así, sería posible sugerir al analista de datos en qué situaciones es suficiente trabajar con el popular modelo binomial negativo o cuándo es mejor recurrir a la familia Poisson-Tweedie. Además, se ilustra la comparación del ajuste de estos modelos sobre el número de consultas pediátricas en un centro de salud de la ciudad de Rosario, Argentina. Si bien los resultados obtenidos fueron similares, se observó una ganancia en la precisión de las estimaciones del modelo Poisson-Tweedie.

***Palabras clave*:** Datos de conteo; Exceso de ceros; Modelos Poisson-Tweedie.

1. Introduction

In studies involving count data, finding an excessive number of zeros is quite common. This excess of zeros is one of the causes of overdispersion (i.e. greater variability than expected). In such cases, the paradigmatic regression model for count data, the Poisson model, turns out to be inappropriate (Hinde & Demétrio, 1998; Zeileis et al., 2008; Agresti, 2015). A very frequent practice is to use the negative binomial model. These models belong to the well-known class of Generalized Linear Models (GLMs) introduced by Nelder & Wedderburn (1972).

Different types of models have been proposed as alternatives for the analysis of data with excessive zeros. Among these, two-part models stand out: hurdle (Mullahy, 1986; Heilbron, 1989) and zero-inflated models (Lambert, 1992; Greene, 1994). Both include an additional linear predictor to describe the excess of zeros. However, precisely due to the complexity derived from the inclusion of an additional predictor, the negative binomial model remains as the most used and popular alternative.

Recently, the proposal developed by Bonat et al. (2018) has been included as another option. This is a new type of model based on the Poisson-Tweedie family (Jørgensen & Kokonendji, 2016), from which the negative binomial model constitutes a particular case. This family can automatically adapt to data characteristics. This is helpful to avoid having to fit a great variety of models and their comparison through tests and goodness-of-fit statistics. Harvey (2020) fitted the Poisson-Tweedie model over a dataset with excessive zeros and found similar

results to those from the well-known negative binomial model. It is important to point out that although this model is covered by the Poisson-Tweedie family, usually it is not fitted into the framework of this family. In practice, the data analyst is used to fit it as a conventional GLM. The similarity found led to the following questions: Are there cases of excessive zeros in which the Poisson-Tweedie family is the best option? Which are they? In other words, our focus is to suggest the data analyst in which situations it would be enough to work with the popular negative binomial model or when it would be best to use the Poisson-Tweedie family. Thus, the aim of this article is to compare the estimates derived from the Poisson-Tweedie model for a wide range of overdispersed data due to excessive zeros with the estimates derived from the negative binomial model.

Section 2 presents a brief review of classical regression models for count data. In Section 3, Poisson-Tweedie models are defined and empirically characterized. Section 4 presents a simulation study to compare the fitting of a Poisson-Tweedie model with that of the negative binomial model. Section 5 illustrates the comparison between these models through a real dataset. The last Section deals with concluding remarks and includes proposals for further research.

2. Classical Models for Count Data

GLMs, introduced by Nelder & Wedderburn (1972), extend the classical linear regression model in order to address non-normal response distributions and non-linear functions of the mean. It is well-known that they are defined by their three components. For a random sample of n observations $(y_i, \mathbf{x}_i)$, where $\mathbf{x}_i$ is the $(q \times 1)$ vector of covariates associated with the $i\text{-}th$ observation $(i = 1, 2, \ldots, n)$, the systematic component of the model, $\mathbf{x}_i{}'\boldsymbol{\beta}$, is related to the expected value of the random component, $E(Y_i|\mathbf{x}_i) = \mu_i$, through a monotonic function called link function $g(.)$, so that

$$g(\mu_i) = \mathbf{x}_i{}'\boldsymbol{\beta} \tag{1}$$

In (1), $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ is the vector of unknown coefficients which are estimated through maximum likelihood using an iterative weighted least squares procedure.

Regarding the random component, the probability distribution of Y belongs to the exponential dispersion family. Thus, the density function for n independent observations y_i is

$$f(y_i; \theta_i; \phi) = a(y_i, \phi) \exp\left\{\frac{y_i\theta_i - \kappa(\theta_i)}{\phi}\right\}$$

where $a(.)$ and $\kappa(.)$ are known functions that determine the considered member of the family (Poisson, binomial, etc.); θ_i is the canonical or natural parameter and ϕ is the dispersion parameter. The g function, for which $g(\mu_i) = \theta_i$, is called canonical or natural link.

In general, the mean and the variance are related through $\kappa''\{\kappa'^{-1}(\mu)\} = v(\mu)$ where $v(.)$ is the so called variance function, which describes the relationship between the mean and the variance.

The simplest model for count data assumes a Poisson distribution for the random component. However, although Zeileis et al. (2008) highlight its usefulness to describe the means μ_i, this model also often underestimates the variance, deriving in liberal tests. This is so since, in many occasions, count observations contradict the equality relationship between the mean and the variance, showing overdispersion.

Another distribution often used is the negative binomial distribution, which can be considered as mixture model by assuming that

$$\begin{aligned} Y \mid \lambda &\sim Poisson(\lambda) \\ \lambda &\sim gamma(\mu, 1/\phi) \end{aligned}$$

with $E(\lambda) = \mu$ and $Var(\lambda) = \phi\mu^2$ (Hinde & Demétrio, 1998; Molenberghs et al., 2007). The expectation and variance of the negative binomial distribution are given by $E(Y) = E(\lambda) = \mu$ and $Var(Y) = E(\lambda) + Var(\lambda) = \mu + \phi\mu^2$.

It can be observed that the greater the dispersion parameter ϕ is, the greater the variance becomes in comparison with the Poisson model variance. For simplicity, the dispersion parameter of the GLMs is considered to be constant for the n observations (Agresti, 2015).

3. Poisson-Tweedie Models

3.1. General Formulation

In order to present the Poisson-Tweedie family, it is necessary to previously define Tweedie models as members of the so called exponential dispersion models. Tweedie densities are characterized by power variance functions of the form

$$v_p(\mu) = \mu^p \quad \text{para} \quad \mu \in \Omega_p \tag{2}$$

where the power parameter $p \in (-\infty, 0] \cup [1, \infty)$ is the index that determines the distribution, $\Omega_0 = \mathbb{R}$, and $\Omega_p = \mathbb{R}_+$ for $p \neq 0$.

The common notation to indicate that a Z variable follows a Tweedie distribution with ϕ and p, dispersion and Tweedie power parameters respectively, is $Z \sim Tw_p(\mu, \phi)$, being μ the mean of Z and the variance $Var(Z) = \phi\mu^p$ for $\mu \in \Omega_p$.

The Tweedie class encompasses known-probability distributions, among which normal distribution ($p = 0$), Poisson distribution ($p = 1$) and Gamma distribution ($p = 2$) can be mentioned. One particular case of interest is the compound Poisson distribution, which corresponds to values of p in the interval $(1, 2)$. This distribution is often chosen to model non-negative data that has a considerable probability mass at zero and is highly right-skewed. This class covers both discrete and continuous distributions, according to the value of the associated power

parameter p. The support of the distribution is given by: the real positive values for $p \geqslant 2$, the non-negative real values for $1 < p < 2$, the natural values for $p = 1$ and the real values for $p \leqslant 0$.

Having the Tweedie models been presented, the Poisson-Tweedie family can be defined by the hierarchical specification

$$\begin{aligned} Y \mid Z &\sim Poisson(Z) \\ Z &\sim Tw_p(\mu, \phi) \end{aligned}$$

in which it is required that the power parameter p be greater than or equal to one, in order to ensure that variable $Z \sim Tw_p(\mu, \phi)$ is non-negative.

The probability mass function of Y for $p > 1$ is given by

$$f(y; \mu, \phi, p) = \int_0^\infty \frac{z^y \exp^{-z}}{y!} a(z, \phi, p) \exp\left\{\frac{z\theta - k_p(\theta)}{\phi}\right\} dz \tag{3}$$

Given the fact that the function $a(z, \phi, p)$ cannot be always written in a closed form, the integral (3) has no simple expression and generally requires recursive algorithms for its calculation. One exception is the case of $p = 2$, which corresponds to a Poisson-Gamma mixture leading to the negative binomial. Despite the difficulty in the integration, the two first moments of the family can be obtained. Jørgensen & Kokonendji (2016) showed that for each $p \geqslant 1$, the Poisson-Tweedie mixture, symbolized as $Y \sim PT_p(\mu, \phi)$, has mean μ, and the variance can be expressed by

$$\begin{aligned} Var(Y) &= \mu + \phi v_p(\mu) \\ &= \mu + \phi \mu^p \end{aligned}$$

where $v_p(\mu)$ is the variance function determined by (2).

When $p = 1$, the integral (3) is replaced by a sum resulting in a Neyman Type A distribution, a mixture Poisson-Poisson (Jørgensen & Kokonendji, 2016; Bonat et al., 2018).

The Tweedie power parameter plays a fundamental role in the Poisson-Tweedie family since it allows capturing overdispersion due to significant skew to the right and an important number of zeros counts. The range of values of p which are related to distributions with a high excess of zeros, according to the results obtained by Bonat et al. (2018), is $(1, 2)$. The value $p = 2$ can be considered as the inflection point between the mentioned distributions and those in which the skewness has a dominant role.

The algebraic structure of the GLMs associated to this family is the same as that of classical GLMs. This paper adopts the log link, although any other function could potentially be considered. For the implementation of these models, it should be noted that the Poisson-Tweedie class is based on assumptions about the moments for estimation and inference. The estimation method resembles Wedderburn's quasi-likelihood method (Wedderburn, 1974) and Liang's generalized estimating equations (Zeger et al., 1988). Therefore, as a full specification of the

probability mass distribution is not available, Bonat et al. (2018) introduce the so-called estimating function approach for the estimation of the Poisson-Tweedie model, where quasi-score estimating functions are adopted for regression parameters and Pearson estimating functions, for parameters ϕ and p (Jørgensen & Knudsen, 2004; Bonat & Jørgensen, 2016). These functions are briefly defined bellow; see Bonat et al. (2018), Bonat & Jørgensen (2016) and Jørgensen & Knudsen (2004) for details.

The quasi-score function for $\boldsymbol{\beta}$ has the following form:

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \begin{bmatrix} \sum_{i=1}^{n} \dfrac{\partial \mu_i}{\partial \beta_1} Var(Y_i)^{-1}(Y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{n} \dfrac{\partial \mu_i}{\partial \beta_q} Var(Y_i)^{-1}(Y_i - \mu_i) \end{bmatrix}$$

where $\partial \mu_i / \partial \beta_j = \mu_i x_{ij}$ for $j = 1, \ldots, q$ and the vector $\boldsymbol{\lambda} = (\phi, p)'$.

The Pearson estimating function for the parameters of the $\boldsymbol{\lambda}$ vector is given by:

$$\psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \begin{bmatrix} -\sum_{i=1}^{n} \dfrac{\partial Var(Y_i)^{-1}}{\partial \phi} \left[(Y_i - \mu_i)^2 - Var(Y_i)\right] \\ -\sum_{i=1}^{n} \dfrac{\partial Var(Y_i)^{-1}}{\partial p} \left[(Y_i - \mu_i)^2 - Var(Y_i)\right] \end{bmatrix}$$

Jørgensen & Knudsen (2004) proposed an iterative algorithm to solve the system of equations

$$\begin{cases} \psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) & = 0 \\ \psi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) & = 0 \end{cases}$$

and showed that the asymptotic distribution of the obtained estimators is normal.

The Poisson-Tweedie model can be easily adjusted in `R` through the `mcglm` package (Bonat, 2016).

3.2. Characterization

Below, the characteristics of several datasets that can be represented by the Poisson-Tweedie family with $p \in (1, 2)$ are shown through a simulation process.

With the purpose of determining scenarios with different dispersion levels and in accordance with Bonat et al. (2018), Fisher's dispersion index (DI) was used. It is defined as

$$DI = \frac{Var(Y)}{E(Y)}$$

It shows how greater the variance is with respect to the mean (Jørgensen & Kokonendji, 2016). DI values considered were 2, 5, 10 and 20, leading to low-dispersion, moderate-dispersion, high-dispersion and very high-dispersion scenarios, respectively. Count values with excessive zeros were randomly generated, following a Poisson-Tweedie distribution $Y \sim PT_p(\mu, \phi)$, with p values equal to 1.1, 1.3, 1.6 and 1.9, μ fixed at 10 and ϕ determined by index DI through the following relationship:

$$DI = 1 + \phi\mu^{p-1} \tag{4}$$

The resulting scenarios allow us to expand the characterization of the Poisson-Tweedie family created by Bonat et al. (2018) within the $(1, 2)$ range, by incorporating the values of p 1.3, 1.6 and 1.9 and keeping the values of DI considered by them.

For the 16 scenarios considered, samples of size 100 000 were generated and empirical probability mass functions were obtained for the respective Poisson-Tweedie distributions. The data generation algorithm was implemented in software `R` (R Core Team, 2019). Initial values were used for generating pseudorandom values so that results reproducibility could be ensured. Function `plot_ ptweedie()` (Dunn, 2013) was used to sketch the graphs of the Poisson-Tweedie densities.

In cases with low dispersion (DI = 2), the shape of graphs is quite similar for the different values of the power parameter. However, when the index DI increases, it can be seen that smaller values of p correspond to situations with more zeros. It can be said that, generally, for $p = 1.1$, overdispersion is fundamentally attributable to an excess of zeros, whereas for $p = 1.9$, apart from having a significant excess of zeros, it is also noticeably skewed to the right, which adds variability (Figure 1).

Another index that makes it possible to explore the flexibility of Poisson-Tweedie distributions is the zero inflated index (ZI). It allows to measure zero inflation and it is given by

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

As it can be seen, the ZI index indicates how many more or how many less zero counts are present in the data in comparison with what is expected under Poisson distribution (assumption by which ZI = 0). In this way, ZI < 0 indicates zero deflation, ZI $= 0$ indicates that there is neither zero deflation nor excess of zeros and ZI > 0 indicates excess of zeros.

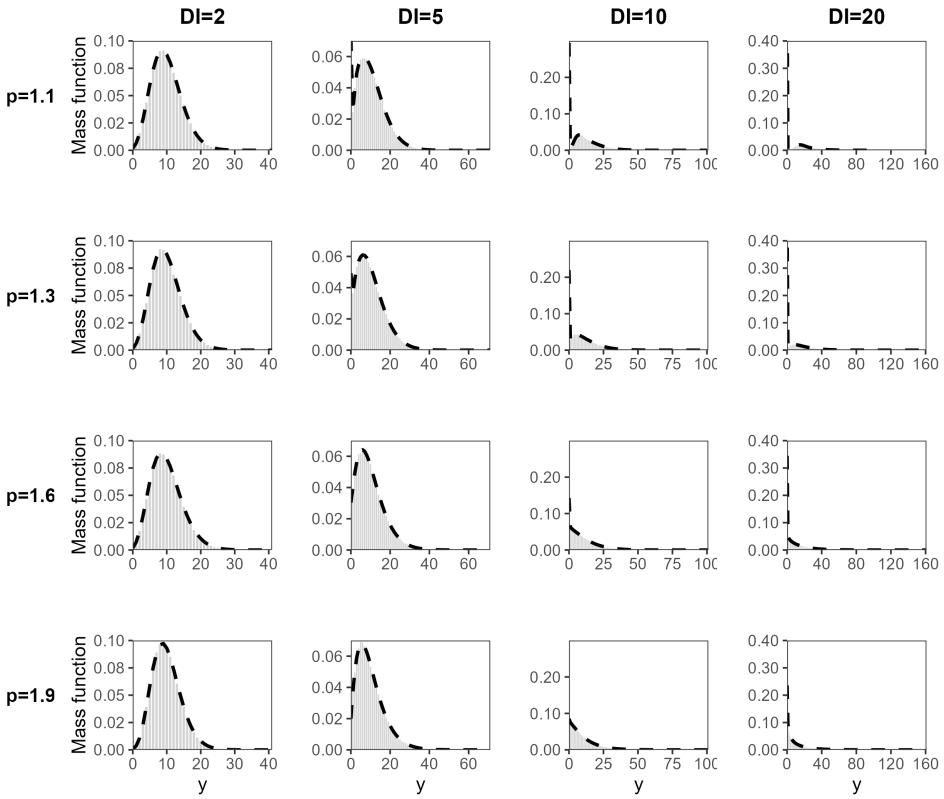

FIGURE 1: Poisson-Tweedie empirical probability mass function (grey) and its approximation by Monte Carlo method (black), according to the values of the dispersion index (DI) and the Tweedie power parameter (p)

It was also worth studying the relationship between both indexes and, in turn, evaluating the potential influence that mean values have on them. For such purposes, the scenarios described above were used, considering, in addition, different values of the parameter μ: 5, 10 and 20. In this case, the `dptweedie()` function (Dunn, 2013) was used to calculate the probability of a zero count according to the different values assumed by distribution parameters.

Figure 2 shows that, for any value of p and μ, the ZI index increases (it gets closer to its upper bound equal to 1) as the DI index increases. In other words, the greater the excess of zero counts is, the greater the overdispersion level present in the data becomes. Furthermore, for the different dispersion levels evaluated, the ZI index decreases with more or less intensity as the value of the power parameter p increases and, in turn, its values are smaller in distributions with greater mean values. That is, the smaller values of p are associated to greater values of ZI, so they represent situations with more excess of zeros. Additionally, it is also found that the relationship between ZI and DI is similar throughout the different mean values considered. Greater ZI values are observed for the smallest value of μ, which

is to be expected since it is assumed that a large excess of zeros corresponds to smaller expected values. Finally, it is worth mentioning that the variation range of ZI is significantly wider for DI = 2, a scenario where even negative values of ZI are observed when $\mu = 20$, indicating zero deflation.

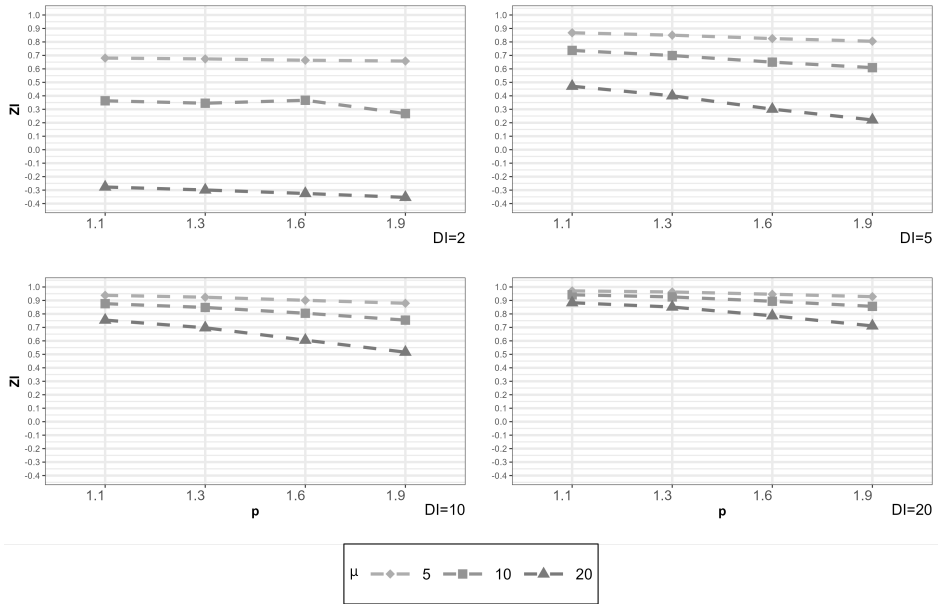

FIGURE 2: ZI index by the values of the Tweedie power parameter (p), the mean value (μ) and the dispersion index (DI)

4. Comparison of the Negative Binomial Model with the Poisson-Tweedie Model

4.1. Study Design

In order to compare the estimations from a negative binomial regression model and Poisson-Tweedie regression model, zero-inflated count data were randomly generated. The algorithm chosen was that proposed by Bonat et al. (2018); the data was simulated under Poisson-Tweedie distribution, $Y_i \sim PT_p(\mu_i, \phi)$, with mean values that comply with the model

$$\begin{aligned} \log(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \\ &= \log(10) + 0.8 x_{1i} - x_{2i} \end{aligned} \tag{5}$$

In (5), μ_i is the mean value of Y for the *i-th* individual, with $i = 1, \ldots, n$; x_{1i} is the value of an equally spaced sequence from -1 to 1 and length equal to the sample size; and x_{2i} is the value of a Bernoulli variable with a probability

equal to 0.5, both assumed by the *i-th* individual. The dispersion parameter, ϕ, was fixed so that the DI index had the following values: 2, 5, 10 and 20, when $\mu_i = 10$, in accordance with (4). The chosen simulation scheme allows to contemplate situations of low, moderate, high and very high-dispersion. According to the characterization presented in Section 3.2, the values $p = 1.1$ and $p = 1.6$ were considered.

In each of the eight scenarios, $m = 1000$ samples of size $n = 100$ were generated and in each of them the negative binomial and Poisson-Tweedie models were fitted. The estimates of the coefficients as well as the estimates of each distribution parameters were obtained. These were: dispersion and power parameter in the Poisson-Tweedie model and dispersion parameter in the negative binomial model and its standard errors.

The estimators behavior was analyzed through the relative percent bias (RB) and the mean squared error (MSE), which are defined as follows:

$$RB = \left[\frac{1}{m}\sum_{i=1}^{m}\hat{\Theta}_i - \Theta\right]\frac{1}{\Theta} \times 100$$

and

$$MSE = (\hat{\Theta}) = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\Theta}_i - \Theta)^2 = Var(\hat{\Theta}) + \left[\frac{1}{m}\sum_{i=1}^{m}\hat{\Theta}_i - \Theta\right]^2$$

where Θ represents the parameter value and $\hat{\Theta}$ is its estimator (Morris et al., 2019).

Confidence intervals of the analyzed coefficients were built considering a nominal confidence level of 95% ($CI_{95\%}$). The coverage rate was calculated as the percentage of $m = 1000$ $CI_{95\%}$ that covered the real coefficient value. The average amplitude of such $CI_{95\%}$ was also calculated.

As to the computational implementation, software `R` (R Core Team, 2019) was used. Just like in the characterization of Poisson-Tweedie models, initial values for the generation of pseudo-random values were used. Function `rptweedie_reg()` (Bonat, 2018) was implemented to generate Poisson-Tweedie data in accordance with the simulation model described. The negative binomial model was fitted by means of function `glm.nb()` of the `MASS` package. The Poisson-Tweedie model was fitted with function `mcglm()`. The conditional standard error of the dispersion parameter was estimated by means of function `mc_conditional_test`. Both functions belong to the `mcglm` package. Finally, calculations of the different evaluation measures of the performance of both models were programmed.

4.2. Results

The estimates obtained for the coefficients of the Poisson-Tweedie and negative binomial models result in the empirical distributions presented in Figures 3, 4 and 5. No differences between the densities corresponding to the estimates deriving from both models are observed. In the different scenarios, densities are centered on the theoretical values with a variability that increases as the DI increases. This can be quantified by analyzing the RB and MSE.

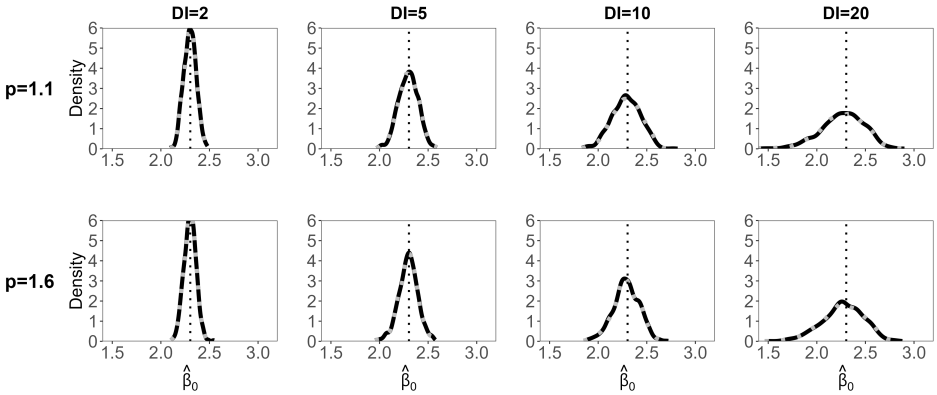

FIGURE 3: Distribution of the coefficient $\beta_0 = 2.3$ estimates in the Poisson-Tweedie model (solid black) and in the negative binomial model (dashed grey) by scenarios, $m = 1000$

In both models, RBs of the three estimated coefficients are very small, representing less than 1% of the value of each of them. Generally, the value of β_0 is underestimated and the values of β_1 and β_2 are overestimated. In practically all scenarios RB increases as DI increases and, in turn, its values are almost always greater in scenarios with $p = 1.1$. It is difficult to identify a clear pattern that systematically favors any of the evaluated models. In some cases, the RB of the Poisson-Tweedie model is equal to that of the negative binomial model, while in other cases, the RB is somehow smaller or greater (Table 1).

TABLE 1: Relative percent bias of the estimated coefficients of the Poisson-Tweedie and negative binomial models by scenarios, $m = 1000$

		DI							
		2		5		10		20	
Coefficient	p	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial
$\beta_0 = 2.3$	1.1	-0.143	-0.139	-0.306	-0.251	-0.557	-0.384	-1.640	-1.844
	1.6	-0.124	-0.115	-0.322	-0.308	-0.584	-0.612	-1.550	-1.357
$\beta_1 = 0.8$	1.1	0.785	0.551	1.113	0.963	2.315	3.489	7.703	0.290
	1.6	0.101	-0.044	-0.079	-0.075	1.643	1.196	3.370	3.542
$\beta_2 = -1.0$	1.1	0.672	0.436	1.409	1.071	3.718	3.887	5.480	-0.643
	1.6	0.496	0.361	0.560	0.488	0.594	0.020	0.647	0.449

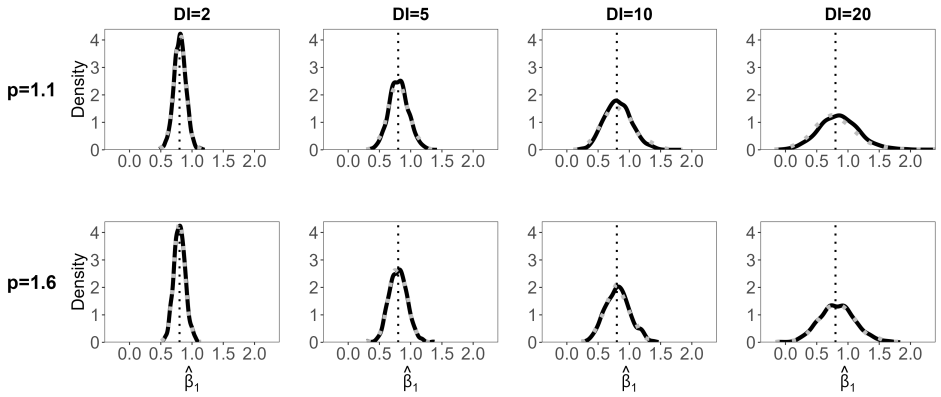

FIGURE 4: Distribution of the regression coefficient $\beta_1 = 0.8$ estimates in the Poisson-Tweedie model (solid black) and in the negative binomial model (dashed grey) by scenarios, $m = 1000$

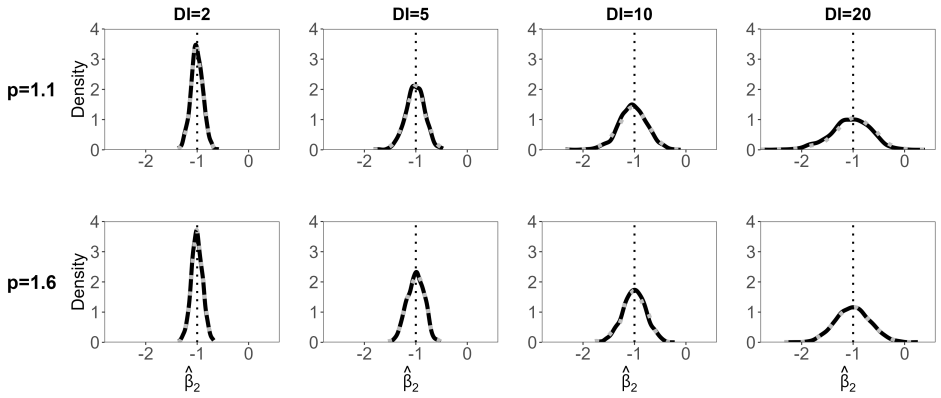

FIGURE 5: Distribution of the regression coefficient $\beta_2 = -1.0$ estimates in the Poisson-Tweedie model (solid black) and in the negative binomial model (dashed grey) by scenarios, $m = 1000$

The analysis of MSEs of the three coefficients shows the same pattern described for the RBs: MSEs are very small in all scenarios and increase as DI increases for the three coefficients. Nevertheless, the Poisson-Tweedie model has a better performance with smaller values, particularly in scenarios with $p = 1.1$ (Table 2).

TABLE 2: Mean squared error of the estimated coefficients of the Poisson-Tweedie and negative binomial models by scenarios, $m = 1000$

		DI							
		2		5		10		20	
Coefficient	p	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial
$\beta_0 = 2.3$	1.1	0.004	0.004	0.010	0.010	0.022	0.023	0.050	0.053
	1.6	0.004	0.004	0.010	0.010	0.018	0.019	0.047	0.047
$\beta_1 = 0.8$	1.1	0.009	0.010	0.024	0.028	0.047	0.067	0.115	0.119
	1.6	0.008	0.008	0.021	0.021	0.039	0.040	0.082	0.083
$\beta_2 = -1.0$	1.1	0.014	0.014	0.035	0.038	0.078	0.091	0.155	0.160
	1.6	0.012	0.012	0.028	0.028	0.051	0.054	0.111	0.117

Concerning the coverage rate of $CI_{95\%}$ of both models, the coverage rate of $CI_{\beta_0;95\%}$ is greater than the nominal value in most scenarios, increasing even up to 98% when DI is equal to 5 and 10. On the contrary, for regression coefficients, coverage rate is, most of the times, lower than the nominal value. In the scenario where there is more dispersion and a large excess of zeros ($p = 1.1$; DI= 20), coverage rates under the negative binomial model are close to 63%, while under the Poisson-Tweedie model, they are close to 93% in all analyzed coefficients. In the remaining scenarios, coverage rates are, in general, slightly lower for the Tweedie-Poisson model (Table 3).

TABLE 3: Coverage rate for the coefficients of the Poisson-Tweedie and negative binomial models by scenarios, $m = 1000$

		DI							
		2		5		10		20	
Coefficient	p	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial
$\beta_0 = 2.3$	1.1	96.2	96.2	95.4	98.8	93.5	98.6	93.9	63.4
	1.6	95.9	95.5	95.2	96.5	94.8	97.8	93.0	97.7
$\beta_1 = 0.8$	1.1	93.7	93.4	93.2	94.4	93.4	95.4	91.4	62.5
	1.6	93.7	94.2	94.2	94.6	93.1	95.4	92.1	96.1
$\beta_2 = -1.0$	1.1	94.7	92.7	93.7	95.1	92.2	95.3	94.1	62.9
	1.6	93.9	94.0	95.1	96.2	93.8	95.4	93.2	96.7

It can be mentioned that the average amplitude of $CI_{95\%}$ is smaller in scenarios with $p = 1.6$ for both models. In addition, regardless of the p parameter value, the comparison between the Poisson-Tweedie model and the negative binomial model shows that the former has more favorable results in scenarios with intermediate dispersion (DI = 5 and DI = 10), whereas in some cases of the remaining scenarios the opposite occurs (Table 4).

It is interesting to mention some issues related to the estimates of the Poisson-Tweedie power parameter and dispersion parameters of both models.

Regarding the estimate of the power parameter, in a significant percentage of samples, the estimates obtained fall outside their parameter space. This occurs especially in the scenarios with $p = 1.1$, in which such percentage varies between 33 and 43%, while when $p = 1.6$, percentages are smaller: they vary between 4 y

TABLE 4: Average amplitude of the $CI_{95\%}$ for the coefficients of the Poisson-Tweedie and negative binomial models by scenarios, $m = 1000$

		DI							
		2		5		10		20	
Coefficient	p	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial	Poisson-Tweedie	Negative binomial
$\beta_0 = 2.3$	1.1	0.257	0.264	0.398	0.521	0.555	0.915	0.798	0.818
	1.6	0.251	0.252	0.386	0.426	0.535	0.645	0.762	1.012
$\beta_1 = 0.8$	1.1	1.012	0.365	0.570	0.666	0.793	1.132	1.131	1.021
	1.6	0.356	0.352	0.531	0.555	0.724	0.812	1.014	1.247
$\beta_2 = -1.0$	1.1	0.459	0.433	0.705	0.779	0.983	1.325	1.409	1.203
	1.6	0.434	0.419	0.637	0.648	0.863	0.948	1.207	1.458

16%. Clearly, fixing scenarios with a power parameter value which is quite close to the border of its parameter space implies a big number of situations presenting estimation problems. In this sense, it should be added that cases in which $\hat{p}$ is smaller than 1 arise because of estimation problems due to non-significant covariates. More specifically, given that the power parameter p is estimated based on the mean-variance relationship, when covariates are not significant, the mean is constant and, therefore, information to estimate such parameter is not available (Bonat, 2018).

Despite the mentioned difficulty, because of the importance attributed to the power parameter when determining the distribution within the family, its behavior is nonetheless described. Figure 6 presents, for each scenario, the empirical densities corresponding to the valid estimates, that is, $\hat{p} > 1$. It can be noted that such estimates are centered on the real value of p (or on a very close value) and that the variability of each scenario is different; scenarios with less dispersion seem to have more variability, whereas scenarios with intermediate dispersion (DI = 5 and DI = 10) show little variability, which increases again when DI = 20. Furthermore, it can also be observed values of $\hat{p}$ that are greater than 2, corresponding to samples highly skewed and that have excessive zeros. In accordance with theoretical values of the power parameter, there are not many cases of this type in scenarios with $p = 1.1$ (they represent less than 2%). On the contrary, in those cases where $p = 1.6$, the percentage is greater, between 5 and 16%. Evidently, these samples are highly right-skewed, a characteristic that prevails over the excess of zeros.

RBs of the p parameter estimate are small for scenarios with $p = 1.6$, which represents a maximum of 3% of such theoretical value, while they increase their value when $p = 1.1$ and DI= 2. In such scenario, they rise up to 18%. In contrast, MSEs are greater in scenarios where $p = 1.6$, which means that an ideal situation where RB and MSE are simultaneously small cannot be found.

Finally, it is important to make some comments about dispersion parameters since difficulties arose when estimating them due to different reasons. As to the Poisson-Tweedie model, because of the existing link between the power parameter and the dispersion parameter ϕ, cases in which the power parameter estimates are outside the parameter space lead to dispersion parameter estimates that are not valid either. In other words, even if the software provides a result, it cannot be considered as a correct value. In the negative binomial model, there are samples

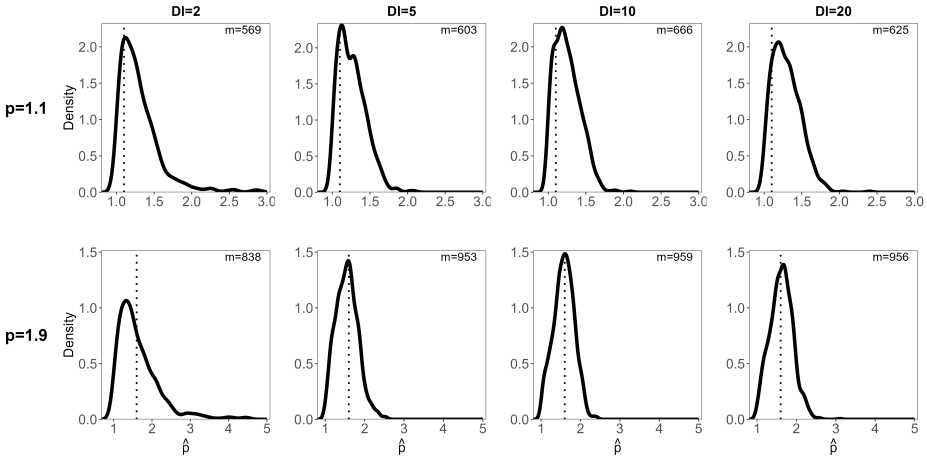

FIGURE 6: Distribution of the power parameter estimates in m samples with $\hat{p} > 1$ by scenarios

where the iterative procedure does not converge and, consequently, the standard error estimates cannot be found. Although in most scenarios such samples do not represent a high percentage, in the scenario that corresponds to the case with the greatest overdispersion and a big number of zero counts ($p = 1.1$; DI= 20), the percentage rises to 56%. In this scenario, in the Poisson-Tweedie model, there are difficulties in a lower percentage: 34% of cases.

5. Illustration

The comparison between the fit of the negative binomial model and that of the Poisson-Tweedie model is illustrated by analysing the number of pediatric visits of 446 children enrolled in a health center in the city of Rosario, Argentina during the year 2019. Data was provided by the Public Health Department of Rosario City Hall.

The purpose is to assess the effect of children's age (between 0 and 4 years old) and the number of visits carried out in the pediatric offices on the demand of consultations in the emergency room of a health center. The marginal distribution of the number of consultations in the emergency room is right-skewed, with a big number of zero counts (44.4%) (Figure 7). Clearly, the dataset has overdispersion: mean equals 1.64 and variance equals 6.04, which is why the use of the Poisson model is discarded from the beginning. Moreover, in this dataset the empirical DI is 3.68. When analyzing conditional distributions, it seems that there is a smaller number of visits in 3 and 4-year-old children, in comparison with those who are younger, as well as a slightly positive correlation between the number of consultations in the pediatric emergency room and in the pediatric office (Figure 8).

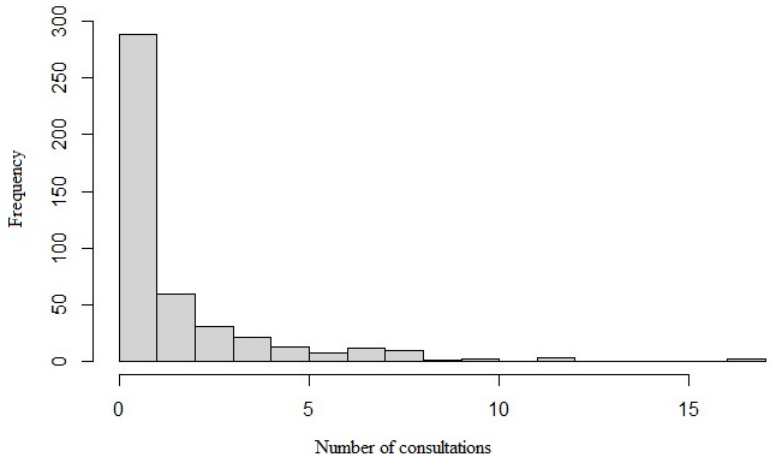

FIGURE 7: Number of consultations in the pediatric emergency room

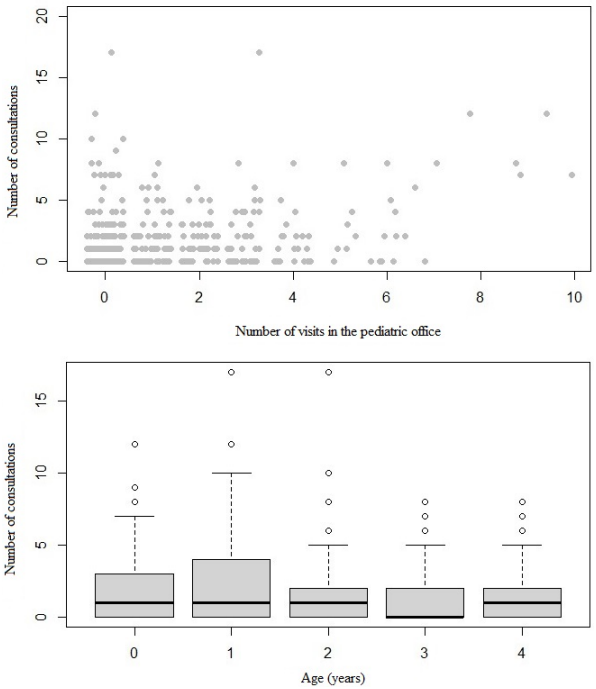

FIGURE 8: Number of consultations in the pediatric emergency room by the number of visits in the pediatric office (above) and children's age (below)

Poisson-Tweedie and negative binomial models are fitted. The number of visits in the pediatric office is included in the linear predictor in its original scale, whereas for children's age, design variables are defined taking as reference children younger than 1 year old. In the Poisson-Tweedie model, the power parameter

estimate falls in the interval $(1, 2)$, being $\hat{p} = 1.525$, as it was expected given the observed excessive zeros. The overdispersion present in the data is evidenced by the significance of the dispersion parameter. Under this model, the mean number of consultations made in the pediatric emergency room is reduced as children are older. It is particularly significant the decrease in the number of consultations when comparing 3 and 4-year-old children with babies younger than 1 year old. In addition, there is a direct association with the number of visits carried out in pediatric offices, regardless of children's age (Table 5).

If instead of resorting to the Poisson-Tweedie family, the classical negative binomial GLM is fitted, the results are very similar regarding both the significance of their effects and their interpretation. It can be observed that estimates become more accurate when searching for the best value of p within the family instead of fitting the negative binomial GLM (Table 5). This results were to be expected based on those found in Section 4.2.

TABLE 5: Parameter estimates and standard errors (SEs) for Poisson-Tweedie and negative binomial models and ratios between them

	Estimates (SEs)		
Parameter	Poisson-Tweedie	Negative binomial	Ratio
Intercept	0.538 (0.186)∗	0.531 (0.193)∗	1.013 (0.964)
Age (1 vs 0)	0.093 (0.211)	0.119 (0.222)	0.782 (0.950)
Age (2 vs 0)	-0.271 (0.234)	-0.257 (0.238)	1.054 (0.983)
Age (3 vs 0)	-0.600 (0.227)∗	-0.588 (0.227)∗	1.020 (1.000)
Age (4 vs 0)	-0.519 (0.238)∗	-0.508 (0.238)∗	1.022 (1.000)
Number of visits	0.129 (0.033)∗	0.125 (0.036)∗	1.032 (0.917)
p	1.525 (0.311)	-	-
ϕ	1.731 (0.413)	1.327 (0.162)	-

∗ Statistically significant at 0.05 level.

6. Concluding Remarks

Upon modeling count data, it is very common to find situations where the number of null results exceeds the expected one in regression models generally used. This work has explored the newest alternative to deal with data with excessive zeros, the Poisson-Tweedie models. It provides a unified framework to handle count data with different characteristics. Special attention has been given to characterize, through a simulation study, Poisson-Tweedie family with power parameter values p which are suitable to capture the excess of zeros. In this sense, the characterization presented by Bonat et al. (2018) was studied in further detail, focusing on the distributions with $p \in (1, 2)$. Among the conclusions obtained, it can be highlighted that the Poisson-Tweedie distribution with $p = 1.1$ represents the situations with the highest excess of zero counts, in accordance with large values of the ZI index. Power parameter values which are close to 2 correspond to situations where not only is the excess of zeros significant, but that also exhibit important skew to the right, adding variability.

In the regression model context, the incorporation of the Poisson-Tweedie model raised the interest to evaluate in which situations it works better than the first chosen model, the negative binomial GLM. The comparison between them was relevant in order to guide the analyst in his daily practice. The simulation study showed similar results in terms of RB, MSE and coverage rate of CIs for the regression model coefficients. Nevertheless, the Poisson-Tweedie model had a better performance when data has zero inflation and high dispersion. It would be interesting to build bootstrap CIs to compare coverage rates and amplitudes without normality assumptions, in future studies.

It is also worth mentioning that in both models, there were problems in the dispersion parameter estimation. They were very noticeable for the negative binomial model when there was strong overdispersion due to a marked excess of zeros (DI = 20 and $p = 1.1$). It was easy to identify the samples with estimations problems in the Poisson-Tweedie power parameter and, therefore, in the dispersion parameter, given their relationship. In contrast, in the negative binomial model, although the dispersion parameter values which are remarkably large introduce doubts about the convergence of the estimation method, it is difficult to find an objective criterion that defines when the estimates are not valid.

As to the application presented, the results obtained in the negative binomial and Poisson-Tweedie models are similar, but the estimates in the latter are more accurate.

Finally, this work has focused on the comparison of the negative binomial model with the family of Poisson-Tweedie models. It remains to be analyzed by means of simulation studies the performance of other models used to handle excessive zeros, such us two-part models and other recent proposals. Specifically, Berger & Tutz (2020) introduce an alternative of a semiparametric class of models used for this type of data, which has the advantage of not requiring assumptions about the count probability distribution. The authors proposed to weigh the advantages of not requiring distributional assumptions against the possibility of choosing the ideal distribution within the great variety of options covered by the Poisson-Tweedie models.

[Received: April 2022 — Accepted: August 2023]

References

Agresti, A. (2015), *Foundations of linear and generalized linear models*, 1 edn, John Wiley & Sons.

Berger, M. & Tutz, G. (2020), *Transition Models for Count Data: A Flexible Alternative to Fixed Distribution Models*. arXiv preprint. arXiv:2003.12411

Bonat, W. H. (2016), *mcglm: Multivariate covariance generalized linear models*. R package version 0.3.0. https://github.com/wbonat/mcglm

Bonat, W. H. (2018), 'Multiple response variables regression models in R: The mcglm package', *Journal of Statistical Software* **84**(4). https://doi.org/10.18637/jss.v084.i04

Bonat, W. H. & Jørgensen, B. (2016), 'Multivariate covariance generalized linear models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(5), 649–675. https://doi.org/10.1111/rssc.12145

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J. & Demétrio, C. G. B. (2018), 'Extended Poisson-Tweedie: Properties and regression models for count data', *Statistical Modelling* **18**(1), 24–49. https://doi.org/10.1177/1471082x17715718

Dunn, P. (2013), *Tweedie exponential family models*. R package version 2.1.7. http://cran.r-project.org/web/packages/tweedie/tweedie

Greene, W. H. (1994), Accounting for excess zeros and sample selection in Poisson and negative binomial regression models, Working Papers 94-10, New York University, Leonard N. Stern School of Business, Department of Economics, New York. https://ideas.repec.org/p/ste/nystbu/94-10.html.

Harvey, G. B. (2020), Estudio de la parasitemia tras la infección por Trypanosoma cruzi en ratas. Ajuste de modelos para datos de conteo con exceso de ceros, Master's thesis, Univesidad Nacional de Rosario.

Heilbron, D. (1989), Generalized linear models for altered zero probabilities and overdispersion in count data, Technical report, Department of Epidemiology and Biostatistics, University of California.

Hinde, J. & Demétrio, C. G. (1998), 'Overdispersion: Models and estimation', *Computational Statistics & Data Analysis* **27**(2), 151–170. https://doi.org/10.1016/s0167-9473(98)00007-3

Jørgensen, B. & Knudsen, S. J. (2004), 'Parameter orthogonality and bias adjustment for estimating functions', *Scandinavian Journal of Statistics* **31**(1), 93–114. https://doi.org/10.1111/j.1467-9469.2004.00375.x

Jørgensen, B. & Kokonendji, C. C. (2016), 'Discrete dispersion models and their tweedie asymptotics', *AStA Advances in Statistical Analysis* **100**(1), 43–78. https://doi.org/10.1007/s10182-015-0250-z

Lambert, D. (1992), 'Zero-inflated Poisson regression, with an application to defects in manufacturing', *Technometrics* **34**(1), 1. https://doi.org/10.2307/1269547

Molenberghs, G., Verbeke, G. & Demétrio, C. G. B. (2007), 'An extended random-effects approach to modeling repeated, overdispersed count data', *Lifetime Data Analysis* **13**(4), 513–531. https://doi.org/10.1007/s10985-007-9064-y

Morris, T. P., White, I. R. & Crowther, M. J. (2019), 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine* **38**(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *Journal of Econometrics* **33**(3), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3

Nelder, J. A. & Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A (General)* **135**(3), 370. https://doi.org/10.2307/2344614

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Wedderburn, R. W. M. (1974), 'Quasi-likelihood functions, generalized linear models, and the gauss-newton method', *Biometrika* **61**(3), 439. https://doi.org/10.2307/2334725

Zeger, S. L., Liang, K.-Y. & Albert, P. S. (1988), 'Models for longitudinal data: A generalized estimating equation approach', *Biometrics* **44**(4), 1049. https://doi.org/10.2307/2531734

Zeileis, A., Kleiber, C. & Jackman, S. (2008), 'Regression models for count data in R', *Journal of Statistical Software* **27**(8). https://doi.org/10.18637/jss.v027.i08