Noname manuscript No.
(will be inserted by the editor)

A new approach to truncated regression for count data

Ana María Martínez-Rodríguez · Antonio Conde-Sánchez · María José Olmo-Jiménez

Received: date / Accepted: date

Abstract Standard Poisson and negative binomial truncated regression models for count data include the regressors in the mean of the non-truncated distribution. In this paper a new approach is proposed so that the explanatory variables determine directly the truncated mean. The main advantage is that the regression coefficients in the new models have a straightforward interpretation as the effect of a change in a covariate on the mean of the response variable. A simulation study has been carried out in order to analyse the performance of the proposed truncated regression models versus the standard ones showing that coefficient estimates are now more accurate in the sense that the standard errors are always lower. Also, the simulation study indicates that the estimates obtained with the standard models are biased. An application to real data illustrates the utility of the introduced truncated models in a hurdle model. Although in the example there are slight differences in the results between the two approaches, the proposed one provides a clear interpretation of the coefficient estimates.

Keywords Count data · Hurdle model · Negative binomial regression · Poisson regression · Truncated models

1 Introduction

Count variables support is typically the set of nonnegative integers. In many cases, however, the entire distribution of counts cannot be observed. In par-

A.M. Martínez-Rodríguez
Department of Statistics and Operations Research, University of Jaén, Paraje Las Lagunillas s/n, 23071 Spain
Tel.: +34-953-212929
E-mail: ammartin@ujaen.es

A. Conde-Sánchez · M.J. Olmo-Jiménez
Department of Statistics and Operations Research, University of Jaén, Paraje Las Lagunillas s/n, 23071 Spain

ticular, the situation in which the zero counts are not recorded, because they are structurally excluded or not observed, is known as zero-truncation. Some examples are: the length of hospital stays, the number of weeks a record is in the top ten list or the number of occupants in a car.

For this case usual models are the Poisson and negative binomial zero-truncated distributions. The regression models developed in this context focus on the mean of the non-truncated distribution (Cameron and Trivedi, 2013; Hilbe, 2011; Winkelmann, 2008). Studies of zero-truncated count data include the length of hospital stay (Hilbe, 2011; Prebensen et al., 2015), the number of recreational visits to a National Park (Martínez-Espiñeira and Amoako-Tuffour, 2008) or the number of cars derailed in a train derailment (Liu et al., 2013), among others. Nevertheless, not only has the zero count not been observed but furthermore the combination of explanatory variables in which the dependent variable becomes zero is unknown. Another important point is that it is not possible to interpret easily the estimated effects since the coefficients in the truncated model do not correspond directly to changes in the truncated mean (Grogger and Carson, 1991). In addition, Hardin and Hilbe (2015) point out that when truncated data are modelled by regression models based on non-truncated distribution, biased estimates can be obtained. Also, they indicate that if the mean of the dependent variable is low (under three of four) then there will be a substantial difference in the coefficient estimated and, as they mention, "despite the closeness of coefficients it is important that we use the appropriate model for the data".

Truncated distributions are also widely used for implementation in hurdle models. Hurdle models (Mullahy, 1986) are used when a separate process seems to generate the positive counts versus the zero counts. In other words, the term hurdle makes reference to a threshold that must be exceeded before events occur, so explanatory variables are allowed to have different influence on each of the two hurdle model components, that is, the factors influencing zeros may be different from those that influence non-zero data (O'Neill and Faddy, 2003). Some examples of this situation are the study about the number of boat trips, where it can be assumed that the decision to participate and the decision about trip frequency can be influenced by different factors (Farr et al., 2014), the non-marital fertility, where the zero count describes the presence of non-marital birth and the positive count is used to explain its intensity (Pazvakawambwa et al, 2014) or the number of publications of a researcher, where once he/she publishes for the first time, he/she passes the hurdle and his/her performance is described by the count component (Baccini et al, 2014).

However, in hurdle models the use of the standard truncated models is a big drawback because as Long and Freese (2014, p. 531) pointed out, it is assumed that the zero and the positive counts come from different data generating processes and the interpretation of the coefficients from the zero-truncated model no longer corresponds directly to changes in the unconditional rate.

For all these reasons it would be interesting to introduce the explanatory variables in the truncated mean so that the regression coefficients could explain

the performance of the true observed mean. In this sense a new approach to truncated regression for count data is proposed in this work. It is along the line of the generalized additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos, 2005, Stasinopoulos and Rigby, 2007) which allow all the parameters of interest (with a clear meaning of location/scale/shape of the distribution of the response) to be modelled as functions of explanatory variables. The Beta regression model (Ferrari and Cribari-Neto, 2004) is also in this direction since the dependent variable is beta-distributed and its mean and its precision parameter are directly related to a set of regressors through a linear predictor.

The paper is structured as follows. Section 2 is dedicated to truncated regression models for count data. Firstly, general truncated distributions are briefly reviewed, including the Poisson and negative binomial, as well as the truncated regression models based on them. Then, the section ends introducing a new methodology for truncated regression models. In Section 3 a simulation study is carried out in order to analyse the performance of the truncated regression models proposed versus the standard ones in relation to the estimates of the coefficients. Section 4 includes an application example of a hurdle model to illustrate the utility of the approach proposed. Finally, Section 5 contains some concluding remarks.

2 Truncated regression models for count data

Let Y be a count variable whose realizations less than a positive integer, say $k+1$, are omitted. Then the resulting distribution is called *left truncated* or *truncated* (for simplicity) and its probability mass function (pmf) is given by

$$f(y|Y>k) = \frac{f(y)}{P\left(Y>k\right)}, \quad y = k+1, k+2, \ldots \tag{1}$$

It is clear that the probabilities in the truncated distribution are greater than in the initial one. Moreover, the following relation between the $r-th$ raw moments is verified (Cameron and Trivedi, 2013)

$$E\left(Y^r|Y>k\right) = \frac{E\left(Y^r\right)}{P(Y>k)} > E\left(Y^r\right). \tag{2}$$

In particular, the truncated mean is greater than the untruncated one. In addition,

$$Var\left(Y|Y>k\right) = E\left(Y|Y>k\right)\left[\frac{Var(Y)}{E(Y)} - E(Y)\cdot\frac{1-P(Y>k)}{P(Y>k)}\right]. \tag{3}$$

Both mean and variance depend on the value of $P\left(Y>k\right)$, so a misspecification of this probability leads to a misspecified truncated mean and to an inconsistent and biased estimation of the untruncated mean (Cameron and Trivedi, 2013; Grogger and Carson, 1991; Santos Silva, 1997).

The most common way of truncation in count models is truncation at 0, that is, $k = 0$, in such a way that the occurrence of an event activates the observation mechanism. Then it is clear that $E(Y|Y > 0) > 1$.

With respect to the Poisson distribution with mean $\mu > 0$, the probability of a zero count is $e^{-\mu}$, so taking (1), (2) and (3) into account

$$\begin{aligned} f(y|Y>0) &= \frac{e^{-\mu}\mu^y}{(1-e^{-\mu})\,y!}, \quad y = 1, 2, \ldots \\ E(Y|Y>0) &= \frac{\mu}{1-e^{-\mu}}, \\ Var(Y|Y>0) &= E(Y|Y>0)\left(1 - \frac{\mu e^{-\mu}}{1-e^{-\mu}}\right). \end{aligned}$$

Let us observe that the truncated variance is less than the truncated mean, so the truncated Poisson (TP) distribution at 0 is always underdispersed.

In the case of the Negative Binomial (NB) with parameters $\mu, \theta > 0$ (where μ is the mean of NB distribution) and pmf[1] given by

$$f(y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\mu}{\theta}\right)^y\left(1+\frac{\mu}{\theta}\right)^{-(y+\theta)}, \quad y = 1, 2, \ldots \tag{4}$$

the probability of a zero count is $(1+\frac{\mu}{\theta})^{-\theta}$. Then, the truncated NB (TNB) distribution at 0 has the following characteristics

$$\begin{aligned} f(y|Y>0) &= \frac{1}{1-(1+\frac{\mu}{\theta})^{-\theta}}\frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)}\left(\frac{\mu}{\theta}\right)^y\left(1+\frac{\mu}{\theta}\right)^{-(y+\theta)}, \quad y = 1, 2, \ldots \\ E(Y|Y>0) &= \frac{\mu}{1-\left(1+\frac{\mu}{\theta}\right)^{-\theta}}, \\ Var(Y|Y>0) &= E(Y|Y>0)\left(1 + \frac{\mu}{\theta} - \frac{\mu(1+\frac{\mu}{\theta})^{-\theta}}{1-(1+\frac{\mu}{\theta})^{-\theta}}\right). \end{aligned}$$

Any of these truncated distributions can be the underlying distribution of a regression model simply considering $\mu = e^{\mathbf{x}'\boldsymbol{\beta}}$, where $\mathbf{x}' = (1 \quad x_1 \quad \cdots \quad x_k)$ is the vector of covariates and $\boldsymbol{\beta}' = (\beta_0 \quad \beta_1 \quad \cdots \quad \beta_k)$ the vector of coefficients (Cameron and Trivedi, 2013). However, these coefficients cannot be directly interpreted in terms of the truncated mean, since the regression is not on this mean but on the untruncated one (O'Neill and Faddy, 2003). In fact, according to (2)

$$\frac{\partial E(Y|\mathbf{x})}{\partial x_j} = \frac{\partial E(Y|Y>0,\mathbf{x})}{\partial x_j}P(Y>0|\mathbf{x}) + E(Y|Y>0,\mathbf{x})\frac{\partial P(Y>0|\,\mathbf{x})}{\partial x_j},$$

so the change in the untruncated mean splits into two components: a part that affects the mean of the currently truncated part of the distribution and a part that affects the probability of truncation. However, it has no sense to

[1] This is the parametrization used in the `R VGAM` package (Yee, 2016).

determine the untruncated mean, that is, to know the effect of the covariates on the untruncated mean given that the zero counts have not actually been observed.

For this reason, we propose to modify the standard truncated regression models for count data so that the covariates affect the truncated mean through a log-linear expression, that is

$$\mu^{tr} = E\left(Y|Y>0\right) = e^{\mathbf{x}'\boldsymbol{\beta}}. \tag{5}$$

Firstly, we shall describe the regression model based on the TP at 0 and next that based on the NB counterpart, though this methodology is applicable to any other regression model with an underlying truncated distribution.

So, if $Y|\mathbf{x}$ follows a TP distribution with mean μ^{tr} we consider

$$\mu^{tr} = \frac{\mu}{1-e^{-\mu}} = e^{\mathbf{x}'\boldsymbol{\beta}}. \tag{6}$$

From now on we note this regression model as $TP\left(\mu^{tr}\right)$.

Then, if $y_1, y_2, \ldots, y_n$ is a sample of size n, the log-likelihood function has the following expression

$$\ln L = \sum_{i=1}^{n} \left\{ y_i \ln(\mu_i) - \mu_i - \ln \Gamma(y_i+1) - \ln\left[1-\exp(-\mu_i)\right] \right\}.$$

Regarding the NB model, if $Y|\mathbf{x}$ follows a TNB distribution with mean μ^{tr} we consider

$$\mu^{tr} = \frac{\mu}{1-\left(1+\frac{\mu}{\theta}\right)^{-\theta}} = e^{\mathbf{x}'\boldsymbol{\beta}}. \tag{7}$$

Similarly, the regression model will be noted $TNB\left(\mu^{tr}\right)$.

If $y_1, y_2, \ldots, y_n$ is a sample of size n, the log-likelihood function is given by

$$\begin{aligned} \ln L = \sum_{i=1}^{n} \Bigg\{ & \ln \Gamma\left(y_i+\theta\right) - \ln \Gamma\left(\theta\right) - \ln \Gamma\left(y_i+1\right) + y_i \ln\left(\frac{\mu_i}{\theta+\mu_i}\right) \\ & + \theta \ln\left(\frac{\theta}{\theta+\mu_i}\right) - \ln\left[1-\left(1+\frac{\mu_i}{\theta}\right)^{-\theta}\right] \Bigg\}. \end{aligned}$$

In both models the estimation of the regression coefficients is carried out by maximizing the respective log-likelihood function using the `optim()` and `nlm()` functions of `R`. In each iteration of the estimation process, the equations given in (6) and (7) for the Poisson and the NB, respectively, have to be solved in order to obtain μ_i. Specifically, the solution of (6) is given by

$$\mu = \mu^{tr} + W\left(-\mu^{tr} e^{-\mu^{tr}}\right). \tag{8}$$

where $W(\cdot)$ is the Lambert $W-$function, also called the omega function or product logarithm (Corless et al, 1996). It is defined as the inverse function of $g(W) = We^{W}$, $W \in \mathbb{C}$ and it is programmed in `R`. The `lambertW()` function

is available in the `R VGAM` package. For its part, the solution of (7) is achieved by numerical methods through the `uniroot()` or `optimize()` functions of `R`.

In addition, in the $TNB\,(\mu^{tr})$ regression model we consider the parametrization $\theta = e^{\theta_0} > 0$ in order to guarantee the positivity of this parameter. Although it is possible to introduce covariates in the latter expression, it exceeds the aim of this work.

It should be pointed out that as (6) or (7) have to be verified and the left-hand-side is always greater than one, then $\mathbf{x}'\boldsymbol{\beta} > 0$.

3 Simulation study

We have carried out a simulation study in order to analyse the performance of the truncated regression models proposed versus the standard ones. Specifically, we have simulated $m = 500$ samples of size $n = 500$ of the $TP(\mu)$, $TP(\mu^{tr})$, $TNB(\mu)$ and $TNB(\mu^{tr})$ regression models and then, we have fitted the standard and the proposed TP or TNB regression models for each sample. We expect that the regression coefficients will be adequately estimated by the regression model that has generated the data.

For the fits we have studied the bias of the coefficient estimates using the mean bias and the standard deviation (s.d.). Furthermore, we have obtained the standard error (s.e.) of the coefficient estimates and the corresponding 95% confidence intervals (CI) for each fit. Next, we have computed the mean of the standard errors and the percentage of those CI that contains the true value of the coefficient, known as coverage.

We have also computed other goodness-of-fit measures, such as the mean squared error (MSE) or the Akaike information criterion (AIC), but they do not show great differences between the proposed regression models and the standard ones so we do not present them.

3.1 Simulation from a zero-truncated Poisson regression model

In this section we have considered two scenarios: in Scenario 1 we have simulated samples of a $TP(\mu)$ regression model and then we have fitted the $TP(\mu)$ and $TP(\mu^{tr})$ regression models for each sample; in Scenario 2, we have simulated samples of a $TP(\mu^{tr})$ regression model and then we have fitted the $TP(\mu)$ and $TP(\mu^{tr})$ regression models for each sample.

The simulation process for Scenario 1 is described as follows:

Step 1. 500 values of two covariates x_1 and x_2 are generated from a uniform distribution on the interval $(0, 1)$.

Step 2. The mean of the Poisson distribution is obtained as $\mu_i = \exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\}$, $i = 1, \ldots, 500$, for several values of the regression coefficients β_0, β_1 and β_2.

Step 3. Each data in the sample is randomly generated from a zero$-TP$ distribution with parameter μ_i using the `trun.r()` function of the `gamlss`

(Stasinopoulos and Rigby, 2016) and `gamlss.tr` (Stasinopoulos and Rigby, 2015) packages of `R`.

Step 4. The coefficients of the $TP(\mu)$ and the $TP(\mu^{tr})$ regression models have been estimated using the `vglm()` function of the `R VGAM` package for the former and the `optimize()` function for the latter.

We have considered values of β_j, $j = 0, 1, 2$, so that μ_i is close to 1 (as in $\beta_0 = 1, \beta_1 = -0.5, \beta_2 = -0.5$) until a maximum value of $e^5 = 148.41316$ (for the case $\beta_0 = 1, \beta_1 = 3, \beta_2 = 1$). We have not taken higher values of the mean into account because, as Hilbe points out (2014a, pp. 175-177), if the mean of the response variable is greater than 5, we simply should not expect any zero counts at all (less than 1% for a Poisson model) and then, there are hardly any differences between a standard and a truncated model.

Simulation results for both regression models and different values of the coefficients are summarized in Table 1. Specifically, Column 1 contains the mean bias and the s.d., in brackets (* indicates a significant bias at 5% level based on a normal 95% CI, given that there are 500 observations). Column 2 shows the mean of the standard errors and Column 3 the percentage of simulations in which the coefficient estimate does not differ significantly from the true value. In the case of the regression model from which data have been generated, this percentage allows us to determine the suitability of the s.e. obtained from the diagonal of the Hessian matrix. With regard to the other regression model this percentage indicates in how many simulations the regression coefficient has been adequately estimated.

From this table we can deduce that, in general, the coefficient estimates for the $TP(\mu)$ regression model are not biased, whereas they are for the $TP(\mu^{tr})$ regression model. Moreover, this bias decreases as the value of the regression coefficient increases, which agrees with the fact that μ and μ^{tr} are very close. In fact, this bias disappears for very high values of the coefficients. It is also noticed that β_0 is overestimated, whereas β_1 and β_2 are underestimated in the $TP(\mu^{tr})$ regression model.

In addition, the means of the coefficient estimate standard errors are lower, in general, for the $TP(\mu^{tr})$ regression model than for the $TP(\mu)$. This seems to show that, despite the bias, the estimates are more accurate.

Regarding the coverage, it approaches 95%, the confidence level considered, in the $TP(\mu)$ regression model, so it shows the validity of the inference made. In the $TP(\mu^{tr})$ regression model, the coverage increases - in general - as the values of the coefficients increase, showing no great differences between the two models.

Among the cases considered we would like to emphasize those with a regression coefficient equal to 0. Our aim is to show the capability of the $TP(\mu^{tr})$ regression model to identify the corresponding non-relevant covariate. As we can observe in the table, the coefficient has been adequately estimated, unlike the rest of the coefficients: a covariate that does not influence the untruncated mean, neither does the truncated mean.

Table 1 Mean bias and s.d. in brackets (* indicates a statistically significant bias at 5% level), average MSE and percentage of 95% CI that contain the true value of the coefficient for Scenario 1

	$TP(\mu)$			$TP(\mu^{tr})$		
	Bias (s.d.)	MSE	Coverage	Bias (s.d.)	MSE	Coverage
$\beta_0 = 1$	0.0005 (0.0990)	0.01907	94	0.0276 (0.0656)*	0.0091	91.4
$\beta_1 = -0.5$	0.0015 (0.1465)	0.0407	94.2	0.1970 (0.0897)*	0.0538	39.2
$\beta_2 = -0.5$	-0.0064 (0.1401)	0.0388	94.8	0.1921 (0.0832)*	0.0508	35.8
$\beta_0 = 1$	-0.0024 (0.0774)	0.0118	94.6	0.0985 (0.0617)*	0.0172	64.6
$\beta_1 = -1$	0.0005 (0.1039)	0.0207	95	0.1988 (0.0802)*	0.0519	28
$\beta_2 = 1$	0.0019 (0.0966)	0.0192	94.6	-0.1966 (0.0758)*	0.0503	25
$\beta_0 = 1$	-0.0078 (0.0712)*	0.0099	94.4	0.0691 (0.0607)*	0.0119	77.4
$\beta_1 = -0.5$	-0.0027 (0.0863)	0.0147	94.2	0.0556 (0.0762)*	0.0145	89.2
$\beta_2 = 1$	0.0121 (0.0910)*	0.0161	94	-0.1061 (0.0783)*	0.0229	69.4
$\beta_0 = 1$	-0.0020 (0.0871)	0.0150	96.4	0.0590 (0.0638)*	0.0116	83.6
$\beta_1 = 0$	-0.0020 (0.1228)	0.0290	94.4	-0.0014 (0.0865)	0.0145	94.2
$\beta_2 = -0.5$	-0.0014 (0.1146)	0.0274	95.6	0.1447 (0.0805)*	0.0344	60.4
$\beta_0 = 1$	-0.0024 (0.0634)	0.0078	95	0.0459 (0.0573)*	0.0085	87.2
$\beta_1 = 0$	-0.0004 (0.0725)	0.0106	95.6	-0.0007 (0.0686)	0.0095	95.6
$\beta_2 = 1$	0.0033 (0.0752)	0.0115	96	-0.0545 (0.0689)*	0.0126	87.8
$\beta_0 = 1$	-0.0019 (0.0741)	0.0110	94.2	0.0657 (0.0617)*	0.0119	82.4
$\beta_1 = 0.1$	-0.0027 (0.0959)	0.0182	93.2	-0.0181 (0.0809)*	0.0132	93.4
$\beta_2 = 0.1$	0.0021 (0.0984)	0.0188	96.8	-0.0139 (0.0830)*	0.0135	94.2
$\beta_0 = 1$	-0.0020 (0.0612)	0.0074	95.4	0.0409 (0.0558)*	0.0078	88.8
$\beta_1 = 0.1$	0.0042 (0.0727)	0.0104	94.6	-0.0009 (0.0693)	0.0094	94.8
$\beta_2 = 1$	0.0010 (0.0740)	0.0109	95.2	-0.0476 (0.0687)*	0.0116	89.8
$\beta_0 = 1$	0.0020 (0.0563)	0.0062	94.8	0.0288 (0.0531)*	0.0063	92.2
$\beta_1 = 0.5$	-0.0026 (0.0646)	0.0083	95.8	-0.0152 (0.0628)*	0.0081	94.4
$\beta_2 = 1$	-0.0006 (0.0666)	0.0088	94.6	-0.0252 (0.0640)*	0.0087	91.6
$\beta_0 = 1$	-0.0000 (0.0479)	0.0043	93.2	0.0094 (0.0465)*	0.0041	92
$\beta_1 = 1.5$	0.0035 (0.0497)	0.0050	95	-0.0045 (0.0489)*	0.0048	95.4
$\beta_2 = 1$	-0.0041 (0.0531)	0.0052	91.8	-0.0097 (0.0524)*	0.0051	91.4
$\beta_0 = 1$	0.0007 (0.0305)	0.0019	96	0.0036 (0.0302)*	0.0019	96.4
$\beta_1 = 3$	-0.0001 (0.0355)	0.0025	93.8	-0.0029 (0.0352)	0.0025	94
$\beta_2 = 1$	-0.0014 (0.0292)	0.0017	95	-0.0025 (0.0291)	0.0017	95

Values of the regression coefficients close to 0 have also been taken into account. In this situation the bias decreases, even being not significant, when another regression coefficient has a high value (for instance $\beta_0 = 1, \beta_1 = 0.1, \beta_2 = 1$) and the coverage approaches 95%. This seems to show that the $TP(\mu^{tr})$ regression model estimates better low coefficients than high ones when both are involved in the expression of the truncated mean.

The simulation process for Scenario 2 is similar to Scenario 1 but now computing the mean of the TP distribution in Step 2 as $\mu_i^{tr} = \exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\}$. Next, the corresponding value of μ_i has to be obtained by the expression given in (8) before going to Step 3.

Table 2 contains the simulation results for Scenario 2 (all the columns have the same meaning as in Table 1). We can infer that, in general, the coefficient estimates for the $TP(\mu)$ regression model are biased, whereas this does not happen for the $TP(\mu^{tr})$ regression model. Again, the bias decreases as the values of the coefficients increase. Moreover, β_0 is underestimated by the $TP(\mu)$ regression model and β_1 and β_2 overestimated. Conclusions for Scenario 2 are similar to those obtained for Scenario 1, but changing roles of the $TP(\mu)$ and $TP(\mu^{tr})$ regression models.

In addition, Figure 1 (and Figure 5 included in "Appendix") contain the boxplots corresponding to the most extreme situations simulated. In each

Table 2 Mean bias and s.d. in brackets (* indicates a statistically significant bias at 5% level), average MSE and percentage of 95% CI that contain the true value of the coefficient for Scenario 2

	$TP(\mu)$			$TP(\mu^{tr})$		
	Bias (s.d.)	MSE	Coverage	Bias (s.d.)	MSE	Coverage
$\beta_0 = 1$	0.1022 (0.1070)*	0.0346	86.2	-0.0034 (0.0593)	0.0069	95.6
$\beta_1 = -0.5$	-0.5916 (0.1779)*	0.4150	8	0.0019 (0.0715)	0.0105	96
$\beta_2 = -0.5$	-0.5898 (0.1880)*	0.4164	9.2	0.0008 (0.0779)	0.0115	93.2
$\beta_0 = 1$	-0.1448 (0.0849)*	0.0351	60.4	0.0006 (0.0613)	0.0075	95.2
$\beta_1 = -1$	-0.2785 (0.1076)*	0.1008	25	-0.0031 (0.0761)	0.0114	94.2
$\beta_2 = 1$	0.2722 (0.1119)*	0.0982	29.6	-0.0008 (0.0797)	0.0120	92
$\beta_0 = 1$	-0.0978 (0.0730)*	0.0200	73.6	-0.0022 (0.0606)	0.0072	94.2
$\beta_1 = -0.5$	-0.0714 (0.0879)*	0.0205	88.2	-0.0031 (0.0752)	0.0113	94.4
$\beta_2 = 1$	0.1483 (0.0904)*	0.0384	61.6	0.0039 (0.0757)	0.0114	94.8
$\beta_0 = 1$	-0.0643 (0.0956)*	0.0224	89.8	-0.0012 (0.0642)	0.0082	94.6
$\beta_1 = 0$	-0.0087 (0.1343)	0.0358	95.2	-0.0056 (0.0839)	0.0141	95.6
$\beta_2 = -0.5$	-0.2875 (0.1253)*	0.1170	44.2	0.0040 (0.0788)	0.0131	95.6
$\beta_0 = 1$	-0.0620 (0.0665)*	0.0122	82.2	-0.0075 (0.0596)*	0.0068	94.4
$\beta_1 = 0$	0.0037 (0.0744)	0.0110	94.8	0.0035 (0.0697)	0.0097	94.8
$\beta_2 = 1$	0.0729 (0.0771)*	0.0172	84.6	0.0075 (0.0702)*	0.0099	95.2
$\beta_0 = 1$	-0.0867 (0.0739)*	0.0189	81.6	-0.0023 (0.0597)	0.0074	96
$\beta_1 = 0.1$	0.0237 (0.0970)*	0.0197	94.8	0.0016 (0.0796)	0.0129	96.8
$\beta_2 = 0.1$	0.0244 (0.0994)*	0.0203	94	0.0022 (0.0816)	0.0132	94.8
$\beta_0 = 1$	-0.0475 (0.0613)*	0.0098	88.2	-0.0005 (0.0558)	0.0062	95.8
$\beta_1 = 0.1$	0.0034 (0.0679)	0.0098	96.6	-0.0016 (0.0648)	0.0089	97
$\beta_2 = 1$	0.0561 (0.0754)*	0.0144	88	0.0024 (0.0695)	0.0095	94.2
$\beta_0 = 1$	-0.0281 (0.0581)*	0.0073	92.2	0.0001 (0.0543)	0.0057	94.8
$\beta_1 = 0.5$	0.0115 (0.0666)*	0.0087	93.8	-0.0015 (0.0648)	0.0081	95.2
$\beta_2 = 1$	0.0270 (0.0671)*	0.0096	93	0.0010 (0.0639)	0.0081	94.6
$\beta_0 = 1$	-0.0138 (0.0412)*	0.0039	96	-0.0046 (0.0401)*	0.0035	97
$\beta_1 = 1.5$	0.0106 (0.0500)*	0.0051	95.2	0.0029 (0.0495)	0.0049	95
$\beta_2 = 1$	0.0085 (0.0478)*	0.0047	96.2	0.0030 (0.0473)	0.0046	95.6
$\beta_0 = 1$	-0.0027 (0.0322)	0.0020	95.2	-0.0001 (0.0319)	0.0020	95.4
$\beta_1 = 3$	0.0031 (0.0342)*	0.0024	95.8	0.0006 (0.0340)	0.0024	95.4
$\beta_2 = 1$	0.0005 (0.0300)	0.0018	94	-0.0004 (0.0300)	0.0018	93.8

graphic the two boxplots on the left refer to Scenario 1 and the other two on the right to Scenario 2. We have used the option `notch` of the `R` function `boxplot()`, which shows the CI for the median as $\pm 1.58 IQR/\sqrt{n}$, where IQR is the InterQuartile Range. This is a nonparametric alternative to the CI for the mean. Those boxplots corroborate the conclusions stated above.

Finally, we would like to point out that the case $\beta_0 = 0$ provides similar results. The differences lie in the fact that the mean is lower than in the case $\beta_0 = 1$. These results are not included for the sake of brevity.

3.2 Simulation from a zero-truncated NB regression model

In the same way as in the previous section, we have considered two scenarios: in Scenario 1, we have simulated samples of a $TNB(\mu)$ regression model and then we have fitted the $TNB(\mu)$ and $TNB(\mu^{tr})$ regression models for each sample; in Scenario 2, we have simulated samples of a $TNB(\mu^{tr})$ regression model and then we have fitted the $TNB(\mu)$ and $TNB(\mu^{tr})$ regression models for each sample.

The simulation process for Scenario 1 is described as follows:

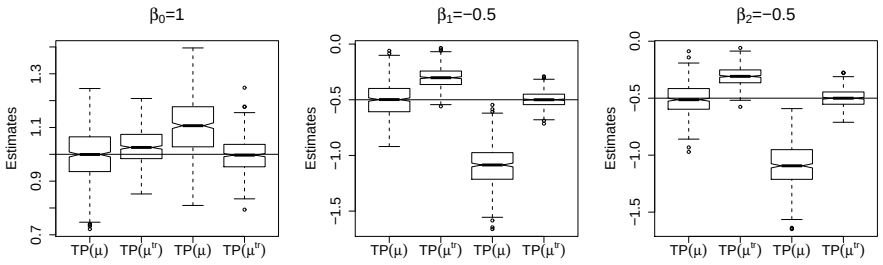

Fig. 1 Boxplots of the coefficient estimates for *TP Scenario 1 (the two on the left in each graphic) and TP Scenario 2 (the two on the right)*

Step 1. 500 values of two covariates x_1 and x_2 are generated from a uniform distribution on the interval $(0, 1)$.

Step 2. The mean of the NB is obtained as $\mu_i = \exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\}$, $i = 1, \ldots, 500$, for several values of the regression coefficients β_0, β_1 and β_2. The value of the parameter θ is fixed.

Step 3. Each data in the sample is randomly generated from a zero$-TNB$ distribution with parameters θ and μ_i using the `trun.r()` function of the `gamlss` and `gamlss.tr` packages of `R`.

Step 4. The coefficients of the $TNB\,(\mu)$ and the $TNB(\mu^{tr})$ regression models have been estimated using the `vglm()` function of the `R VGAM` package for the first one and the `optimize()` function for the second.

To begin with, we have chosen low values for the θ parameter since if $\theta \to \infty$ the NB distribution converges to the Poisson and there are no significant differences between the two models. Specifically, we have considered $\theta = 1, 1/2$ and $1/3$ because, as Hilbe points out (2011, p. 190), the value of $\alpha = 1/\theta$ rarely is greater than 4. However, let us remember that, in practice, we use $\theta_0 = \ln \theta$.

With regard to the regression coefficients, we have considered greater values than in the TP regression model since the lower is θ, the more zero counts the NB distribution has and the more differences there are between the NB and the TNB distributions. Increasing the values of β_1 and β_2 also increases the mean and decreases the differences between both distributions.

For reasons of brevity, only results for $\theta = 1$ and $\theta = 1/3$, that is, $\theta_0 = 0$ and $\theta_0 = -\ln 3 = -1.0986$ are included.

For $\theta = 1$ (Table 3) the same as mentioned for the Poisson can be seen: the coefficient estimates for the $TNB\,(\mu)$ regression model are unbiased, whereas they are biased for the $TNB\,(\mu^{tr})$ regression model. This bias decreases as the value of the coefficients increases. Nevertheless, the bias is significant even for these values, with the only exception of $\beta_j = 0$, which is appropriately estimated, just as in the Poisson. Moreover, the coefficient estimates are better when the true values are low (in absolute terms) than when there is a greater value, although the bias does not disappear.

Table 3 Average bias and s.d. in brackets (* indicates a statistically significant bias at 5% level), average of the coefficient estimate MSEs and percentage of 95% CI that contain the true value of the coefficient for Scenario 1 and $\theta = 1$

	$TNB(\mu)$			$TNB\left(\mu^{tr}\right)$		
	Bias (s.d.)	MSE	Coverage	Bias (s.d.)	MSE	Coverage
$\theta_0 = 0$	0.0245 (0.1738)	0.0633	96	0.0221 (0.1740)	0.0633	96
$\beta_0 = 1$	-0.0119 (0.1536)	0.0458	93.6	0.3201 (0.1021)*	0.1232	11.6
$\beta_1 = -1$	0.0231 (0.1923)*	0.0711	93.6	0.2974 (0.1375)*	0.1242	39.2
$\beta_2 = 1$	-0.0050 (0.1847)*	0.0678	95.8	-0.2854 (0.1307)*	0.1153	39.2
$\theta_0 = 0$	0.0108 (0.1666)	0.0534	93.2	0.0102 (0.1671)	0.0536	93.6
$\beta_0 = 1$	0.0099 (0.1451)	0.0419	93.2	0.3256 (0.1043)*	0.1277	13.2
$\beta_1 = -0.5$	-0.0091 (0.1791)	0.0635	95.4	0.1081 (0.1371)*	0.0490	86.4
$\beta_2 = 1$	-0.0114 (0.1854)	0.0662	94.6	-0.2357 (0.1408)*	0.0938	59.2
$\theta_0 = 0$	0.0306 (0.2194)*	0.0959	93.6	0.0303 (0.2195)*	0.0960	93.8
$\beta_0 = 1$	0.0024 (0.1577)	0.0491	94.6	0.3090 (0.0971)*	0.1145	10.4
$\beta_1 = 0$	-0.0086 (0.1850)	0.0695	96	-0.0057 (0.1252)	0.0319	96
$\beta_2 = -0.5$	-0.0014(0.1859)	0.0700	95.2	0.1595 (0.1266)*	0.0575	75.6
$\theta_0 = 0$	0.0215 (0.1426)*	0.0409	94.6	0.0215 (0.1426)*	0.0409	94.6
$\beta_0 = 1$	-0.0022 (0.1352)	0.0376	94.8	0.2961 (0.1030)*	0.1093	19.2
$\beta_1 = 0$	-0.0006 (0.1632)	0.0560	96.4	-0.0003 (0.1330)	0.0372	96.2
$\beta_2 = 1$	-0.0035 (0.1727)	0.0596	96	-0.1871 (0.1397)*	0.0738	70.8
$\theta_0 = 0$	0.0086 (0.0785)*	0.0123	95.6	0.0066 (0.0786)	0.0123	95.4
$\beta_0 = 1$	-0.0020 (0.1195)	0.0303	95.2	0.1955 (0.1046)*	0.0614	59
$\beta_1 = 0$	-0.0077 (0.1599)	0.0512	94	-0.0087 (0.1508)	0.0453	93.8
$\beta_2 = 5$	0.0068 (0.1601)	0.0526	95.2	-0.2475 (0.1427)*	0.1024	59
$\theta_0 = 0$	0.0290 (0.1823)*	0.0687	97	0.0291 (0.1823)*	0.0687	97
$\beta_0 = 1$	-0.0089 (0.1484)	0.0442	95.6	0.3046 (0.1011)*	0.1132	13.4
$\beta_1 = -0.1$	0.0140 (0.1783)	0.0645	94	0.0369 (0.1309)*	0.0359	93.8
$\beta_2 = 0.1$	0.0066 (0.1858)	0.0671	93.8	-0.0221 (0.1360)*	0.0365	94
$\theta_0 = 0$	0.0138 (0.1372)*	0.0382	92.8	0.0136 (0.1371)	0.0382	92.8
$\beta_0 = 1$	0.0049 (0.1375)	0.0382	95	0.3002 (0.1049)*	0.1122	17.8
$\beta_1 = 0.1$	-0.0107 (0.1769)	0.0606	94.4	-0.0274 (0.1453)*	0.0416	94.4
$\beta_2 = 1$	-0.0022 (0.1665)	0.0573	95.6	-0.1790 (0.1365)*	0.0702	76.8
$\theta_0 = 0$	0.0173 (0.1307)*	0.0333	94	0.0172 (0.1307)*	0.0333	94
$\beta_0 = 1$	0.0044 (0.1361)	0.0370	95.2	0.2827 (0.1081)*	0.1029	24
$\beta_1 = 0.5$	-0.0071 (0.1718)	0.0578	93.4	-0.0819 (0.1448)*	0.0479	90.4
$\beta_2 = 1$	-0.0000 (0.1643)	0.0555	96.8	-0.1505 (0.1396)*	0.0622	81.2
$\theta_0 = 0$	0.0121 (0.0754)*	0.0111	92.2	0.0111 (0.0755)*	0.0110	92.2
$\beta_0 = 1$	0.0006 (0.1293)	0.0322	94.2	0.1494 (0.1151)*	0.0479	70.8
$\beta_1 = 5$	-0.00314 (0.1550)	0.0498	96.6	-0.1668 (0.1433)*	0.0702	79
$\beta_2 = 1$	-0.0092 (0.1717)	0.0545	93	-0.0535 (0.1639)*	0.0526	71
$\theta_0 = 0$	0.0077 (0.0605)*	0.0076	94.8	0.0071 (0.0605)*	0.0076	94.6
$\beta_0 = 1$	0.0063 (0.1212)	0.0294	93.8	0.0750 (0.1150)*	0.0318	90.2
$\beta_1 = 5$	-0.0080 (0.1540)	0.0482	95	-0.0620 (0.1504)*	0.0497	93.2
$\beta_2 = 5$	-0.0079 (0.1569)	0.0492	95.6	-0.0620 (0.1537)*	0.0508	93.2

In addition, the means of the standard errors are lower for the $TNB(\mu^{tr})$ regression model than for the $TNB(\mu)$ and the coverage is around 95% for the right model.

Concerning the θ_0 parameter, it is - in general - slightly underestimated with significant bias in some cases. Taking into account that this parameter is fixed, it should have been estimated in a similar way by both modellings. In fact, the coverage is next to 95%, so it seems to be appropriately estimated.

The image for $\theta = 1/3$ (Table 4) is similar but augmented compared with those obtained for $\theta = 1$, that is, the bias is greater for the same regression coefficients. This is due to the fact that the NB distribution has more zero counts as θ decreases and so there are more differences in relation to the TNB. The most significant change is in the estimates of θ_0 which now are unbiased in some cases.

The simulation process for Scenario 2 is similar to Scenario 1 but taking into account that the mean of the TNB is now obtained in Step 2 with the expression $\mu_i^{tr} = \exp\{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}\}$, $i = 1, \ldots, 500$. Then, before going to Step 3, the corresponding value of μ_i has to be obtained solving the equation given in (7) by means of the `R uniroot()` function.

Table 4 Average bias and s.d. in brackets (* indicates a statistically significant bias at 5% level), average of the coefficient estimate MSEs and percentage of 95% CI that contain the true value of the coefficient for Scenario 1 and $\theta = 1/3$

	$TNB(\mu)$			$TNB\left(\mu^{tr}\right)$		
	Bias (s.d.)	MSE	Coverage	Bias (s.d.)	MSE	Coverage
$\theta_0 = -1.0986$	0.0143 (0.2523)*	1.3100	94.2	0.0128(0.2528)*	1.3138	94.4
$\beta_0 = 1$	-0.0053(0.2349)	0.1128	95	0.6581 (0.1324)*	0.4687	0.2
$\beta_1 = -1$	0.0149 (0.2481)	0.1204	93.8	0.2901 (0.1762)*	0.1453	63.4
$\beta_2 = 1$	-0.0107 (0.2382)	0.1155	94.2	-0.2866 (0.1698)*	0.1411	62
$\theta_0 = -1.0986$	0.0048 (0.2442)	1.3159	95.8	0.0040 (0.2433)	1.3173	96
$\beta_0 = 1$	-0.0187 (0.2429)	0.1135	95.6	0.6437 (0.1472)*	0.4549	0.6
$\beta_1 = -0.5$	-0.0045 (0.2499)	0.1207	94.4	0.1198 (0.1896)*	0.0833	90
$\beta_2 = 1$	0.0220 (0.2449)*	0.1190	95.2	-0.2278 (0.1830)*	0.1184	74.2
$\theta_0 = -1.0986$	0.0182 (0.2832)	1.3382	93.6	0.0181(0.2832)	1.3388	93.6
$\beta_0 = 1$	-0.0042 (0.2610)	0.1333	93.4	0.6455 (0.1403)*	0.4530	0.2
$\beta_1 = 0$	-0.0131 (0.2513)	0.1215	95.2	-0.0088 (0.1747)	0.0588	94.4
$\beta_2 = -0.5$	0.0042 (0.2389)	0.1153	94.8	0.1554 (0.1662)*	0.0797	84.8
$\theta_0 = -1.0986$	0.0132 (0.2314)	1.2812	94.4	0.0131 (0.2315)	1.2815	94.4
$\beta_0 = 1$	-0.0057 (0.2168)	0.0975	95.8	0.6381 (0.1326)*	0.4448	0.2
$\beta_1 = 0$	0.0032 (0.2316)	0.1118	94.4	0.0030 (0.1809)	0.0685	94.4
$\beta_2 = 1$	-0.0023 (0.2321)	0.1124	96	-0.2188 (0.1801)*	0.1158	79.2
$\theta_0 = -1.0986$	0.0157 (0.1311)*	1.2055	93.8	0.0140 (0.1313)*	1.2095	94.2
$\beta_0 = 1$	0.0011 (0.2025)	0.0809	95.2	0.5519 (0.1623)*	0.3557	6.2
$\beta_1 = 0$	-0.0058 (0.2541)	0.1265	95.2	-0.0064 (0.2258)	0.0993	95.6
$\beta_2 = 5$	-0.0081 (0.2656)	0.1345	93.4	-0.5396 (0.2248)*	0.3868	28.8
$\theta_0 = -1.0986$	0.0142 (0.2816)	1.3308	92.6	0.0142 (0.2816)	1.3309	93
$\beta_0 = 1$	0.0102 (0.2367)	0.1164	95.6	0.6605 (0.1322)*	0.4717	0.2
$\beta_1 = -0.1$	-0.0295 (0.2420)*	0.1171	93.2	0.0056 (0.1766)	0.0621	94.8
$\beta_2 = 0.1$	-0.0105 (0.2412)	0.1159	94.4	-0.0342 (0.1762)*	0.0630	93.4
$\theta_0 = -1.0986$	0.0159 (0.2127)	1.2646	94.2	0.0158 (0.2129)	1.2650	94.4
$\beta_0 = 1$	-0.0218 (0.2188)*	0.0977	95.8	0.6217 (0.1389)*	0.4258	0.8
$\beta_1 = 0.1$	0.0252 (0.2212)*	0.1075	97.2	-0.0018 (0.1740)	0.0662	97.2
$\beta_2 = 1$	-0.0024 (0.2435)	0.1178	94.6	-0.2147 (0.1908)*	0.1184	79
$\theta_0 = -1.0986$	0.0093 (0.2103)	1.2725	94.4	0.0091 (0.2103)	1.2730	94.8
$\beta_0 = 1$	0.0029 (0.2242)	0.0979	95	0.6332 (0.1435)*	0.4422	0.2
$\beta_1 = 0.5$	-0.0046 (0.2534)	0.1228	93.6	-0.1022 (0.2036)*	0.0899	90.2
$\beta_2 = 1$	-0.0128 (0.2428)	0.1181	95.2	-0.2049 (0.1947)*	0.1178	82
$\theta_0 = -1.0986$	0.0142 (0.1137)*	1.2020	94.4	0.0129 (0.1137)*	1.2048	94.2
$\beta_0 = 1$	-0.0107 (0.1893)	0.0757	95.6	0.4929 (0.1550)*	0.2927	14
$\beta_1 = 5$	0.0028 (0.2382)	0.1212	96.8	-0.4304 (0.2135)*	0.2799	52
$\beta_2 = 1$	0.0006 (0.2488)	0.1252	95	-0.0998 (0.2252)*	0.1125	92.4
$\theta_0 = -1.0986$	0.0097 (0.0883)*	1.2013	95	0.0086 (0.0883)*	1.2037	94.8
$\beta_0 = 1$	0.0020 (0.1904)	0.0750	95.4	0.3758 (0.1729)*	0.1996	38.8
$\beta_1 = 5$	-0.0136 (0.2539)	0.1313	95.6	-0.2552 (0.2410)*	0.1817	82.8
$\beta_2 = 5$	-0.0049 (0.2561)	0.1320	94.8	-0.2464 (0.2434)*	0.1786	82.2

Simulation results for Scenario 2 are in Tables 3 and 4 for $\theta = 1$ and $\theta = 1/3$, respectively. The conclusions obtained from the simulation of the zero$-TNB$ regression models are the same as from the simulation of the zero$-TP$ regression models.

Finally, boxplots for some of these simulations are shown in Figures 2, 3, 4 (and Figure 6 included in "Appendix"). In each graphic the two boxplots on the left refer to Scenario 1 and the other two on the right to Scenario 2. These graphs confirm the above-mentioned conclusions.

3.3 Conclusions of the simulation study

To sum up we can say that in Scenario 1 the standard zero-truncated regression models fit adequately, but this is not the case for the proposed regression models in which the estimates are biased. Specifically, β_0 is overestimated whereas β_1 and β_2 are underestimated.

In Scenario 2 the role of the two regression models changes. Now, the proposed regression models are the most appropriate: the standard regression

Table 5 Average bias and s.d. in brackets (* indicates a statistically significant bias at 5% level), average of the coefficient estimate MSEs and percentage of 95% CI that contain the true value of the coefficient for Scenario 2 and $\theta = 1$

	$TNB(\mu)$			$TNB\left(\mu^{tr}\right)$		
	Bias (s.d.)	MSE	Coverage	Bias (s.d.)	MSE	Coverage
$\theta_0 = 0$	0.0021 (0.2339)	0.1077	94.6	0.0309 (0.2313)	0.1071	94.4
$\beta_0 = 1$	-0.5603 (0.1753)*	0.3750	7.6	-0.0074 (0.0932)	0.0170	95.2
$\beta_1 = -1$	-0.6512 (0.2067)*	0.5098	11.2	0.0029 (0.1145)	0.0255	93.8
$\beta_2 = 1$	0.6694 (0.2146)*	0.5375	9.4	0.0084 (0.1144)	0.0256	93.2
$\theta_0 = 0$	0.0268 (0.1904)*	0.0730	94.2	0.0292 (0.1903)	0.0732	94.2
$\beta_0 = 1$	-0.4771 (0.1539)*	0.2762	11	-0.0058 (0.0940)	0.0183	95.8
$\beta_1 = -0.5$	-0.2071 (0.1874)*	0.1126	79	0.0052 (0.1286)	0.0329	94.4
$\beta_2 = 1$	0.4279 (0.1858)*	0.2533	36.8	0.0047 (0.1253)	0.0320	95
$\theta_0 = 0$	0.0247 (0.3248)	0.2186	97.4	0.0249 (0.3249)	0.2187	96.8
$\beta_0 = 1$	-0.4296 (0.2029)*	0.2658	40.2	0.0036 (0.0890)	0.0156	96.2
$\beta_1 = 0$	-0.0003 (0.2071)*	0.0890	95	-0.0000 (0.1074)	0.0239	94.6
$\beta_2 = -0.5$	-0.4722 (0.2261)*	0.3217	44.6	-0.0084 (0.1156)	0.0257	94.6
$\theta_0 = 0$	0.0233 (0.1631)*	0.0528	94.4	0.0236 (0.1634)*	0.0530	94.6
$\beta_0 = 1$	-0.4210 (0.1499)*	0.2217	17.8	-0.0048 (0.1012)	0.0204	95
$\beta_1 = 0$	0.0028 (0.1678)	0.0593	96	0.0020 (0.1278)	0.0346	96.2
$\beta_2 = 1$	0.2939 (0.1920)*	0.1553	62.8	-0.0026 (0.1454)	0.0392	93
$\theta_0 = 0$	0.0105 (0.0819)*	0.0130	93.2	0.0093 (0.0816)*	0.0129	93.2
$\beta_0 = 1$	-0.2281 (0.1286)*	0.0850	56	-0.0090 (0.1091)	0.0240	95.8
$\beta_1 = 0$	0.0040 (0.1611)	0.0518	95.6	0.0034 (0.1468)	0.0436	95
$\beta_2 = 5$	0.2894 (0.1647)*	0.1383	59.2	0.0053 (0.1447)	0.0420	95.2
$\theta_0 = 0$	0.0521 (0.2534)*	0.1286	94.8	0.0520 (0.2536)*	0.1289	94.6
$\beta_0 = 1$	-0.4564 (0.1700)*	0.2664	50	-0.0031 (0.0948)	0.0178	94.6
$\beta_1 = -0.1$	-0.0582 (0.1887)*	0.0765	94.6	-0.0002 (0.1195)	0.0293	95.6
$\beta_2 = 0.1$	0.0608 (0.1955)*	0.0795	93	0.0019 (0.1235)	0.0303	96
$\theta_0 = 0$	0.0142 (0.1614)*	0.0504	94.8	0.0143 (0.1614)*	0.0505	94.6
$\beta_0 = 1$	-0.4082 (0.1479)*	0.2103	19	-0.0008 (0.1015)	0.0207	94.8
$\beta_1 = 0.1$	0.0202 (0.1772)*	0.0627	94.6	-0.0061 (0.1366)	0.0374	95.2
$\beta_2 = 1$	0.2781 (0.1720)*	0.1386	66.6	-0.0009 (0.1323)	0.0360	95.2
$\theta_0 = 0$	0.0082 (0.1430)	0.0395	94.6	0.0082 (0.1429)	0.0395	94.6
$\beta_0 = 1$	-0.3792 (0.1455)*	0.1854	23	-0.0056 (0.1024)	0.0212	95.4
$\beta_1 = 0.5$	0.1146 (0.1731)*	0.0727	88.6	0.0020 (0.1396)	0.0390	95
$\beta_2 = 1$	0.2280 (0.1610)*	0.1079	73.8	0.0059 (0.1285)	0.0360	97.4
$\theta_0 = 0$	0.0122 (0.0704)*	0.0104	93.8	0.0111 (0.0704)*	0.0104	93.8
$\beta_0 = 1$	-0.1632 (0.1283)*	0.0587	75.4	-0.0067 (0.1132)	0.0253	94.2
$\beta_1 = 5$	0.1702 (0.1609)*	0.0808	83	-0.0031 (0.1468)	0.0438	95.4
$\beta_2 = 1$	0.0555 (0.1617)*	0.0542	93	0.0099 (0.1538)	0.0465	93.8
$\theta_0 = 0$	0.0066 (0.0605)*	0.0076	95	0.0057 (0.0605)*	0.0076	94.8
$\beta_0 = 1$	-0.0703 (0.1199)*	0.0340	92	-0.0041 (0.1136)	0.0262	94.2
$\beta_1 = 5$	0.0495 (0.1565)*	0.0514	93.6	-0.0021 (0.1527)	0.0468	93.6
$\beta_2 = 5$	0.0520 (0.1507)*	0.0500	94.8	-0.0002 (0.1469)	0.0451	95.8

model underestimates the parameter β_0 and overestimates the parameters β_1 and β_2.

Nevertheless, the coefficient estimates are more accurate in the proposed regression models than in the standard ones, since the s.e. are always lower.

Focusing on the TP regression model, there are no differences in the coefficient estimates as the mean increases. This performance is not the same for the TNB regression model, since the bias is lower when the mean increases but it does not disappear.

4 Application example

The most common application of truncated models is in hurdle models. A hurdle model (Mullahy, 1986; Cameron and Trivedi, 2013) is a modified count model in which there are two processes, one generating the zero counts and one generating the positive counts. The two models are not constrained to be the same. The concept underlying the hurdle model is that a binomial probability model governs the binary outcome of whether a count variable has a zero

Table 6 Average bias and s.d. in brackets (* indicates a statistically significant bias at 5% level), average of the coefficient estimate MSEs and percentage of 95% CI that contain the true value of the coefficient for Scenario 2 and $\theta = 1/3$

	$TNB(\mu)$			$TNB\left(\mu^{tr}\right)$		
	Bias (s.d.)	MSE	Coverage	Bias (s.d.)	MSE	Coverage
$\theta_0 = -1.0986$	-0.0369 (0.4177)*	1.6744	94	0.0073 (0.4044)*	1.5515	93.6
$\beta_0 = 1$	-1.1188 (0.3472)*	1.5172	1.8	0.0008 (0.10734	0.0226	94.6
$\beta_1 = -1$	-0.8451 (0.2673)*	0.8557	9	0.0006 (0.1356)	0.0350	93
$\beta_2 = 1$	0.8300 (0.2682)*	0.8311	10.4	-0.0083 (0.1356)	0.0351	93
$\theta_0 = -1.0986$	-0.0018 (0.3332)	1.4510	96.6	0.0021 (0.3296)	1.4391	96.8
$\beta_0 = 1$	-1.0341 (0.3011)*	1.2547	1.4	-0.0072 (0.1220)	0.0283	92.4
$\beta_1 = -0.5$	-0.2746 (0.2483)*	0.1980	80.4	0.0128 (0.1545)	0.0474	93.6
$\beta_2 = 1$	0.5688 (0.2456)*	0.4465	34.8	-0.0080 (0.1474)	0.0451	96.6
$\theta_0 = -1.0986$	-0.0179 (0.5872)	4.1322	93.4	-0.0175 (0.5849)	2.0322	93.4
$\beta_0 = 1$	-1.0182 (0.4535)*	3.5926	24.8	-0.0036 (0.0994)	0.0198	93.6
$\beta_1 = 0$	-0.0017 (0.2571)	0.1347	95.4	-0.0021 (0.1231)	0.0307	95.2
$\beta_2 = -0.5$	-0.5356 (0.2597)*	0.4245	48.8	0.0012 (0.1208)	0.0301	95.4
$\theta_0 = -1.0986$	0.0462 (0.2785)*	1.2676	92.6	0.0464 (0.2783)*	1.2671	92.8
$\beta_0 = 1$	-0.9420 (0.2514)*	1.0192	0.8	-0.0087 (0.1214)	0.0297	94.8
$\beta_1 = 0$	0.0276 (0.2411)*	0.1166	95.4	0.0203 (0.1657)*	0.0553	95.8
$\beta_2 = 1$	0.4242 (0.2576)*	0.3054	59.2	-0.0076 (0.1758)	0.0585	92.8
$\theta_0 = -1.0986$	0.0189 (0.1237)*	1.1983	96	0.0217 (0.1224)*	1.1920	95.8
$\beta_0 = 1$	-0.7036 (0.1937)*	0.5737	4.6	-0.0035 (0.1368)	0.0405	96.2
$\beta_1 = 0$	0.0124 (0.2395)	0.1189	95.4	0.0095 (0.1988)	0.0834	95.2
$\beta_2 = 5$	0.7133 (0.2546)*	0.6382	20.2	-0.0103 (0.2044)	0.0842	95.4
$\theta_0 = -1.0986$	0.0243 (0.4349)	1.5993	94.4	0.0241 (0.4346)	1.5972	94.4
$\beta_0 = 1$	-0.9972 (0.3277)*	1.2800	3.6	0.0009 (0.1043)	0.0229	95.2
$\beta_1 = -0.1$	-0.0666 (0.2489)*	0.1275	94	0.0040 (0.1449)	0.0416	94
$\beta_2 = 0.1$	0.0574 (0.2417)*	0.1229	94.2	-0.0084 (0.1396)	0.0401	95.4
$\theta_0 = -1.0986$	0.0177 (0.2609)	1.3169	96	0.0180 (0.2607)	1.3161	96
$\beta_0 = 1$	-0.9460 (0.2483)*	1.0252	0.8	-0.0067 (0.1196)	0.0296	95.6
$\beta_1 = 0.1$	0.0396 (0.2460)*	0.1203	94.4	-0.0018 (0.1725)	0.0583	95.4
$\beta_2 = 1$	0.4228 (0.2397)*	0.2959	58.2	0.0025 (0.1666)	0.0563	94.4
$\theta_0 = -1.0986$	0.0186 (0.2646)	1.3016	94	0.0188 (0.2639)	1.3007	93.8
$\beta_0 = 1$	-0.9017 (0.2637)*	0.9441	1.8	-0.0031 (0.1341)	0.0342	93.8
$\beta_1 = 0.5$	0.1877 (0.2465)*	0.1541	88.4	0.0057 (0.1784)	0.0630	93.4
$\beta_2 = 1$	0.3410 (0.2485)*	0.2370	70.6	-0.0096 (0.1801)	0.0638	93.4
$\theta_0 = -1.0986$	0.0104 (0.1161)*	1.2119	95.2	0.0110 (0.1154)*	1.2105	95.4
$\beta_0 = 1$	-0.6125 (0.2042)*	0.4574	12.4	-0.0035 (0.1546)	0.0480	95.2
$\beta_1 = 5$	0.5319 (0.2584)*	0.4144	45	-0.0163 (0.2166)	0.0947	95
$\beta_2 = 1$	0.1460 (0.2576)*	0.1506	90.8	0.0182 (0.2274)	0.1012	95
$\theta_0 = -1.0986$	0.0042 (0.0911)	1.2140	94.2	0.0038 (0.0909)	1.2148	94.2
$\beta_0 = 1$	-0.4055 (0.1999)*	0.2432	44	0.0023 (0.1667)	0.0565	96
$\beta_1 = 5$	0.2653 (0.2537)*	0.2015	83.4	-0.0018 (0.2351)	0.1133	95.2
$\beta_2 = 5$	0.2485 (0.2603)*	0.1960	83.2	-0.0168 (0.2396)	0.1160	95.8

or a positive value. If the value is positive, the "hurdle is crossed" and the conditional distribution of the positive values is governed by a zero-truncated count model. Formally,

$$P(Y = y) = \begin{cases} f_1(0) & y = 0 \\ (1 - f_1(0))\, f_2(y) & y > 0 \end{cases}$$

where $f_1(0) = P(Y = 0)$ and $f_2(y)$ is the pmf of the corresponding zero-truncated distribution.

We shall illustrate the new approach introduced using a NB hurdle regression model. Specifically, we use data from Fair (1978) about the number of marital affairs reported by the respondent within the last year, available in the *affairs* data set of the `R COUNT` package (Hilbe, 2014b). Observed values of this variable appear in Table 9.

Taking into account that these data show a clear overdispersion ($mean = 1.455907$ and $variance = 10.8818$), Hilbe (2014a) has modelled them, using a NB regression model. Specifically, he has considered the following covariates:

– *kids*: 1 if the couple have children and 0 if not.

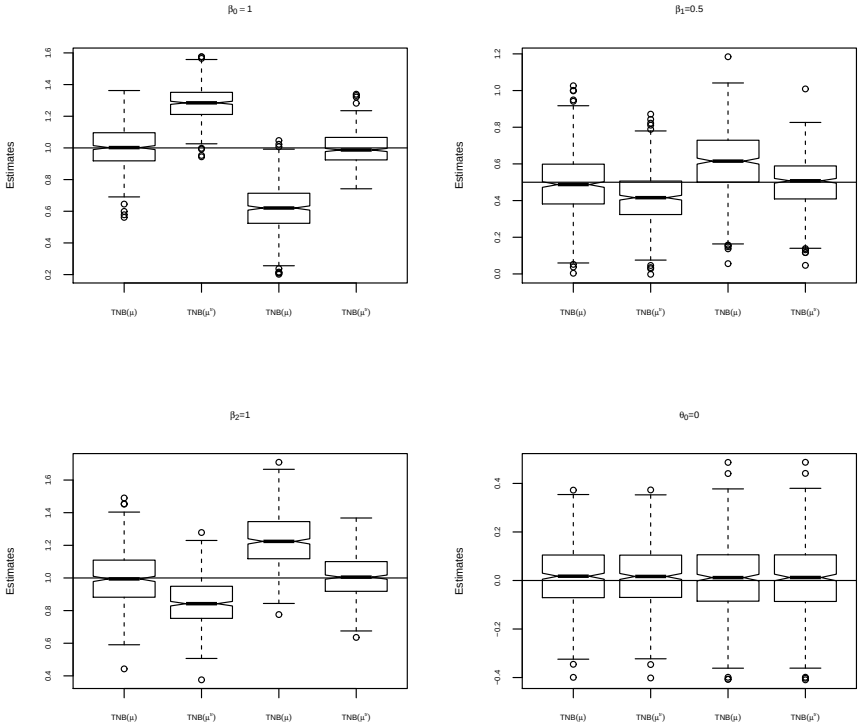

Fig. 2 Boxplots of the coefficient estimates for *TNB Scenario 1 (the two on the left in each graphic) and TNB Scenario 2 (the two on the right)*

- *how rate marriage* with four categories: very unhappy or somewhat unhappy, average (*avgmarr*), happier than average (*hapavg*) and very happy (*vryhap*). The first is the reference category.
- *how religious* with 5 categories: antireligious, not religious (*notrel*), slightly religious (*slghtrel*), somewhat religious (*smerel*) and very religious (*vryrel*). The first is the reference category.
- *years of marriage* with 6 categories: less than 3 years, from 3 to 5 years (*yrsmarr3*), from 6 to 8 years (*yrsmarr4*), from 9 to 11 years (*yrsmarr5*) and 12 or more years (*yrsmarr6*). The first is the reference category.

Nevertheless, these data are susceptible of being modelled by means of a hurdle model since the process generating zeros may be different from that generating the positive counts, that is, there is a zero barrier or hurdle: once a person has succumbed to the temptation of being unfaithful, there are other factors that can influence the number of times he or she is unfaithful again.

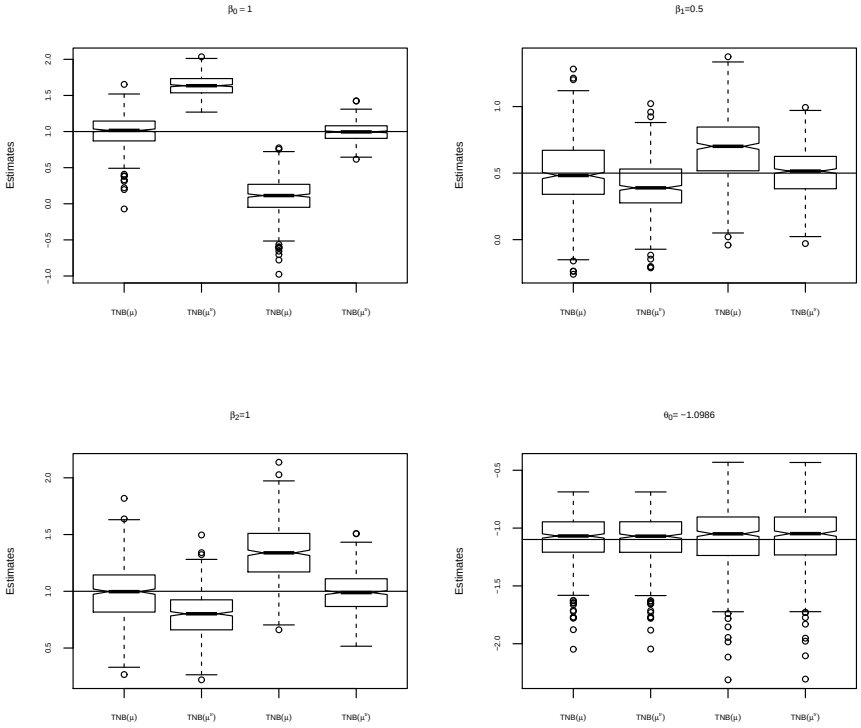

Fig. 3 Boxplots of the coefficient estimates for *TNB Scenario 1 (the two on the left in each graphic) and TNB Scenario 2 (the two on the right)*

In the example, we first fit a hurdle regression model using a logit model for the zero counts, that is

$$f(0|\mathbf{x}) = P(Y = 0|\mathbf{x}) = \frac{1}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}} \tag{9}$$

and a zero-truncated NB regression model for the positive counts. We have used the function `hurdle()` of the `R` `pscl` package (Jackman, 2015). Results are summarized in Table 7.

Secondly, we carry out the fit considering the NB hurdle model with covariates in the truncated mean. Results are summarized in Table 8 (only those corresponding to the positive counts are shown since the logit model coefficients are the same).

This fit is similar to the standard one since the value of the AIC is almost the same, but the most interesting is that the parameter estimates are different, their s.e. are lower and the significant variables change: only the variables $yrsmarr4$, $yrsmarr5$ and $yrsmarr6$ are statistically significant at 5% level. This implies that the only factor that influences the number of marital affairs

Table 7 Negative binomial-logit hurdle regression (* indicates statistically significant coefficient at 5% level)

	Estimate	s.e.	Statistic	$p-$value
Zero counts				
(Intercept)	0.1080	0.4775	0.226	0.8210
kids	0.1817	0.3101	0.586	0.5580
avgmarr	-0.8281	0.3324	-2.491	0.0127*
hapavg	-1.1355	0.2904	-3.910	0.0001*
vryhap	-1.6067	0.3109	-5.168	0.0000*
notrel	-0.9880	0.3698	-2.671	0.0076*
slghtrel	-0.5934	0.3694	-1.606	0.1082
smerel	-1.5175	0.3745	-4.052	0.0001*
vryrel	-1.4244	0.4549	-3.131	0.0017*
yrsmarr3	0.7172	0.3808	1.883	0.0597
yrsmarr4	0.7433	0.4287	1.734	0.0829
yrsmarr5	1.0644	0.4333	2.457	0.0140*
yrsmarr6	0.8892	0.3878	2.293	0.0219*
Positive counts				
(Intercept)	1.7729	0.3967	4.469	0.0000*
kids	-0.2573	0.2205	-1.167	0.2432
avgmarr	-0.4820	0.2347	-2.054	0.0400*
hapavg	-0.3817	0.1935	-1.973	0.0485*
vryhap	-0.4367	0.2262	-1.931	0.0535
notrel	0.0984	0.2627	0.375	0.7080
slghtrel	-0.1057	0.2578	-0.410	0.6819
smerel	-0.4543	0.2706	-1.679	0.0932
vryrel	-0.5698	0.3437	-1.658	0.0974
yrsmarr3	0.0433	0.3101	0.140	0.8890
yrsmarr4	0.5394	0.3228	1.671	0.0948
yrsmarr5	0.5695	0.3158	1.804	0.0713
yrsmarr6	0.7812	0.2882	2.711	0.0067*
$\theta_0 = \ln(\theta)$	-0.5846	0.2519		
$AIC = 1433.222$				

Table 8 Negative binomial-logit hurdle regression with covariates in the truncated mean (* indicates statistically significant coefficient at 5% level)

	Estimate	s.e.	Statistic	$p-$value
Positive counts				
(Intercept)	1.5625	0.3552	4.3993	0.0000*
kids	-0.2124	0.1935	-1.0976	0.2724
avgmarr	-0.3413	0.2124	-1.6074	0.1080
hapavg	-0.2847	0.1775	-1.6039	0.1087
vryhap	-0.2986	0.2051	-1.4558	0.1454
notrel	0.2340	0.2378	0.9840	0.3251
slghtrel	0.0386	0.2352	0.1643	0.8695
smerel	-0.2891	0.2411	-1.1995	0.2303
vryrel	-0.3982	0.3088	-1.2896	0.1972
yrsmarr3	0.1935	0.2500	0.7741	0.4389
yrsmarr4	0.6397	0.2751	2.3256	0.0200*
yrsmarr5	0.6383	0.2627	2.4297	0.0151*
yrsmarr6	0.8417	0.2378	3.5394	0.0004*
$\theta_0 = \ln(\theta)$	-0.4444	0.2619		
$AIC = 1433.424$				

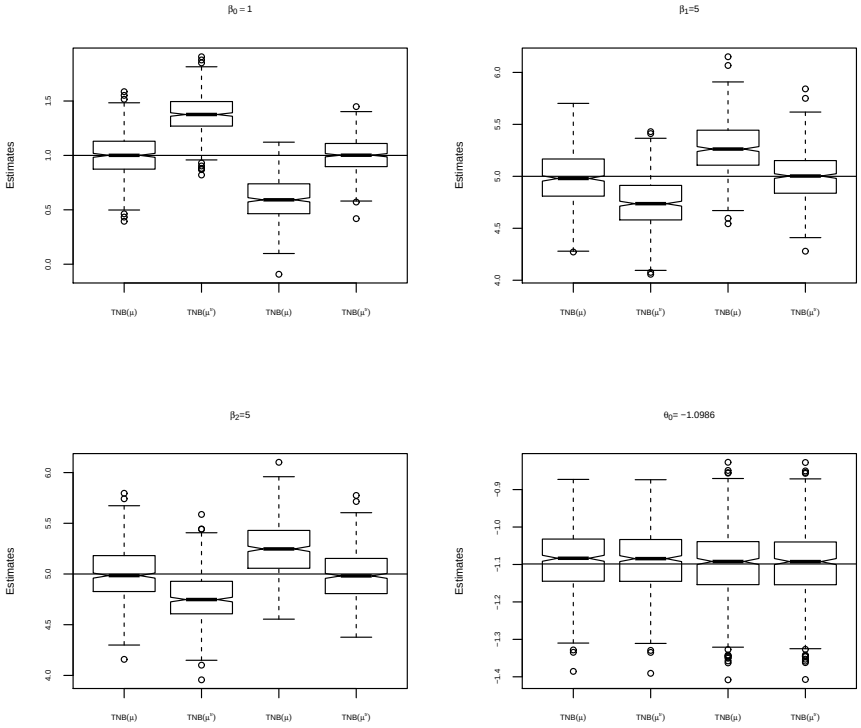

Fig. 4 Boxplots of the coefficient estimates for *TNB Scenario 1 (the two on the left in each graphic) and TNB Scenario 2 (the two on the right)*

is the number of years married. In fact, this number increases as the time of marriage increases. Specifically, the number of affairs of a couple who has been married for more than 5 years and less than 9, increases on average by $e^{0.6397} = 1.8959$ affairs, in relation to another couple, in the same conditions, who have been married for less than 3 years. This average becomes $e^{0.6383} = 1.8933$ affairs when the first couple has been married for more than 8 years and less than 12 and $e^{0.8417} = 2.3203$ affairs if the time of marriage exceeds 11 years.

Let us emphasize that these interpretations are not possible in the first NB hurdle regression model fitted since the coefficients in the truncated model do not correspond directly to changes in the truncated mean.

Another difference between the two fitted hurdle regression models, is the θ estimate: 0.6412 in the proposed model and 0.5573 in the standard one.

Regarding the logit model that generates the zero counts, we can deduce that being happier implies crossing the hurdle with less probability. The same

happens with how religious you are. However, the more years a couple has been married, the more likely they are to be unfaithful.

Finally, Table 9 contains the observed and expected frequencies for each hurdle regression model fitted (the standard and the proposed ones).

In short, the fits are similar, the coefficient estimates and their s.e. change (in fact, the latter decrease) but now these coefficients can be interpreted.

Table 9 Observed and expected frequencies about the number of marital affairs using the standard and the proposed hurdle regression models

		Expected	
y	Observed	$TNB(\mu)$	$TNB(\mu^{tr})$
0	451	451	451
1	34	22	24
2	17	21	21
3	19	19	18
4	0	16	15
5	0	13	13
6	0	11	10
7	42	9	8
8	0	7	7
9	0	6	6
10	0	5	5
11	0	4	4
12	38	3	3
≥ 13	0	14	16

5 Conclusions

In this paper, a new approach to truncated regression for count data including the regressors in the mean of the truncated distribution has been proposed. We have focused on Poisson and NB distributions although other authors have used different truncated distributions (Gurmu, 1998; Gonzales-Barron et al, 2014; Chou et al, 2015; Hardin and Hilbe, 2015) but in all the cases the regressors explain the mean of the non-truncated distribution. The main advantage of the proposed approach is that, although in goodness of fit similar results have been obtained, the parameters in the new models have a straightforward interpretation as the effect of a change in a covariate on the response variable.

In addition, the new approach is very interesting in hurdle models where the mechanism generating the positive counts differs from the one generating zeros and so the interpretation of the coefficients from the zero-truncated model does not correspond to changes of the non-truncated mean. Even more, the observed counts reveal information about the mechanism that generates the positive counts and not about the non-truncated distribution, and it is essential to use the most suitable model for the data.

Acknowledgements We are grateful to the anonymous referees for their careful review and constructive comments that substantially improved the article.

References

Baccini, A., Barabesi, L., Cioni, M. and Pisani, C., Crossing the hurdle: the determinants of individual scientific performance. Scienctometrics, **101 (3)**, 2035-2062 (2014)

Cameron, A. C. and Trivedi, P. K., Regression Analysis of Count Data. Cambridge University Press, New York (2013)

Chou, Y., Chuang, H. H. and Shao B. B. M., Information initiatives of mobile retailers: a regression analysis of zero-truncated count data with underdispersion. Applied Stochastic Models in Business and Industry, 31, 457-463 (2015)

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J. and Knuth, D. E., On the Lambert $W-$function. Advances in Computational Mathematics, 5(4), 329-359 (1996)

Fair, R., A Theory of Extramarital Affairs. Journal of Political Economy, 86: 45-61 (1978)

Farr, M., Stoeckl, N. and Sutton, S., Recreational fishing and boating: Are the determinants the same? Marine Policy, 47, 126-137 (2014)

Ferrari, S. L. P., Cribari-Nieto, F., Beta regresssion for modelling rates and proportions, Journal of Applied Statistics, 31(7),799-815 (2004)

Gonzales-Barron, U., Cadavez, V. and Butler, F., Conducting inferential statistics for low microbial counts in foods using the Poisson-gamma regression. Food Control, 37, 385-394 (2014)

Grogger, J.T. and Carson, R.T., Models for truncated counts, Journal of Applied Econometrics, 6, 225-238 (1991)

Gurmu, S., Generalized hurdle count data regression models, Economics Letters, 58, 263-268 (1998)

Hardin, J. W. and Hilbe, J. M., Regression models for count data from truncated distributions, The Stata Journal, 15, 226-246 (2015)

Hilbe, J. M., Negative Binomial Regression. Cambridge University Press, New York (2011)

Hilbe, J. M., Modeling Count Data. Cambridge University Press, New York (2014a)

Hilbe, J. M., Functions, data and code for count data, R package version 1.3.2, URL http://CRAN.R-project.org/package=COUNT (2014b)

Jackman, S., pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University, R package version 1.4.9, URL http://CRAN.R-project.org/package=pscl, (2015)

Liu, X., Saat, M. R., Qin, X. and Barkan, C. P. L., Analysis of U.S. freight-train derailment severity using zero-truncated negative binomial regression and quantile regression, Accident Analysis and Prevention, 5, 87-93 (2013)

Long, J. S. and Freese, J., Regression Models for Categorical Dependent Variables Using Stata. Stata Press (2014)

Martínez-Espiñeira, R. and Amoako-Tuffour, J., Recreation demand analysis under truncation, overdispersion, and endogenous stratification: An application to Gros Morne National Park, Journal of Environmental Management, 88, 1320-1332 (2008)

Mullahy, J., Specification and testing of some modified count data models, Journal of Econometrics, 33, 341-365 (1986)

O'Neill, M. F. and Faddy, M. J., Use of binary and truncated negative binomial modelling in the analysis of recreational catch data, Fisheries Research, 60, 471-477 (2003)

Pazvakawambwa, L, Indongo, N. and Kazembe, L., A hurdle negative binomial regression model for non-marital fertility in Namibia, Journal of Mathematics and System Science, 4, 498-508 (2014)

Prebensen, N. K., Altin, M. and Uysal, M., Length of stay: A case of Northerm Norway, Scandinavian Journal of Hospitality and Tourism, 15 (Supplement 1), 28-47 (2015)

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

Santos Silva, J. M. C., Generalized Poisson regression for positive count data, Communications in Statistics - Simulation and Computation, 26(3), 1089-1102 (1997)

Stasinopoulos D. M. and Rigby R. A., Generalized additive models for location, scale and shape (GAMLSS) in R, Journal of Statistical Software, 23(7), 1-46 (2007)
Stasinopoulos D. M. and Rigby R. A., Generating and fitting truncated (gamlss.family) distributions, R package version 4.3-1, URL http://CRAN.R-project.org/package=gamlss.tr (2015)
Stasinopoulos D. M. and Rigby R. A., Generalised Additive Models for Location Scale and Shape, R package version 4.4-0, URL http://CRAN.R-project.org/package=gamlss (2016)
Winkelmann, R., Econometric analysis of count data. Springer-Verlag, Heidelberg (2008)
Yee, T. W., VGAM: Vector Generalized Linear and Additive Models, R package version 1.0-1, URL http://CRAN.R-project.org/package=VGAM (2016)

Appendix

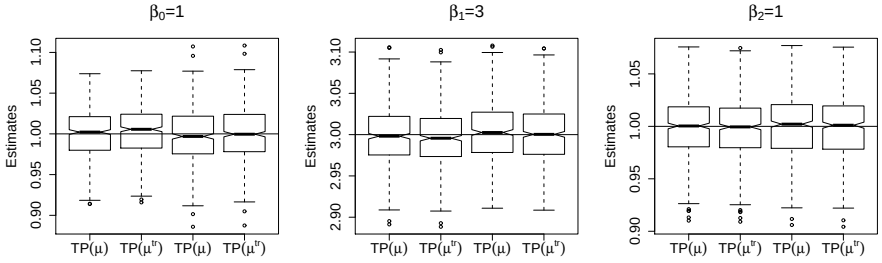

Fig. 5 Boxplots of the coefficient estimates for *TP Scenario 1 (the two on the left in each graphic) and TP Scenario 2 (the two on the right)*

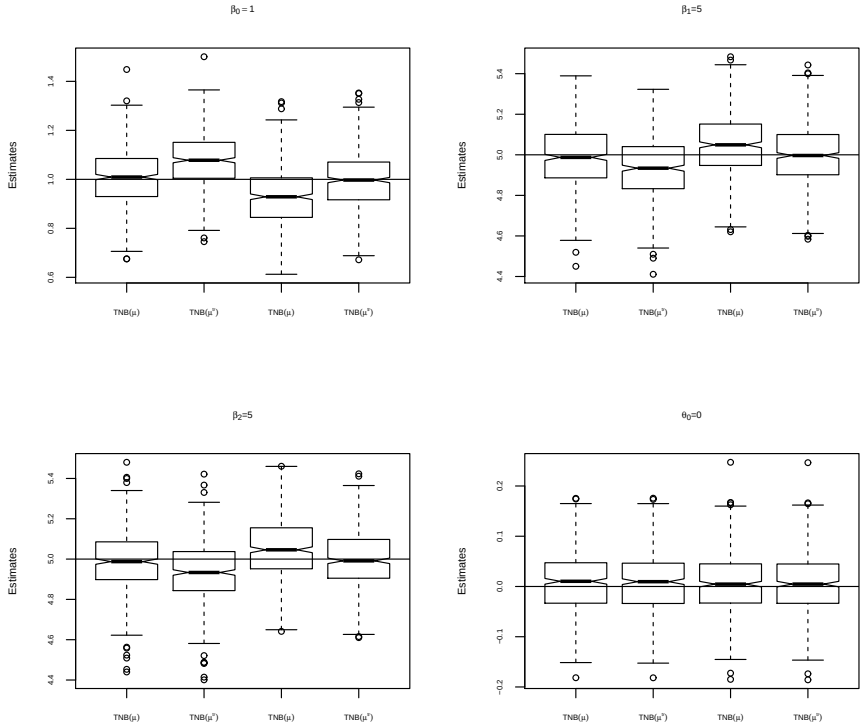

Fig. 6 Boxplots of the coefficient estimates for *TNB Scenario 1 (the two on the left in each graphic) and TNB Scenario 2 (the two on the right)*