Austrian Journal of Statistics
January 2022, Volume 51, 28–39.
http://www.ajs.or.at/
doi:10.17713/ajs.v51i2.1227

Zero-inflated Modified Borel-Tanner Regression Model for Count Data

Anwar Hassan
Dep. of Statistics
University of Kashmir

Ishfaq S. Ahmad
Dep. of Statistics
University of Kashmir

Peer Bilal Ahmad
Dep. of Mathematical Sciences
IUST

Abstract

By starting from the one-parameter Modified Borel-Tanner distribution proposed recently in the statistic literature, we introduce the zero-inflated Modified Borel-Tanner distribution. Additionally, on the basis of the proposed zero-inflated distribution, a novel zero-inflated regression model is proposed, which is quite simple and may be an interesting alternative to usual zero-inflated regression models for count data. The parameters of the proposed model are estimated by Maximum Likelihood Estimation technique. To check the potentiality of the zero inflated Modified Borel-Tanner regression, an application to the count of infected blood cells is taken. The results suggest that the new zero inflated Modified Borel-Tanner regression is more appropriate to model these count data than other familiar zero-inflated (or not) regression models commonly used in practice.

Keywords: MBT distribution, count data, excess zeros, over-dispersion, zero-inflated models.

1. Introduction

Without any ambiguity, Poisson model is one of the basic and simplest count data model and most common in practice to deal with count data. The Poisson model assumes that the events taken into consideration occurs under the principle of complete randomness, but this principle always does not hold true. As we are already aware that the Poisson distribution is characterized by one parameter, has its mean equal to the variance. As the mean and variance of the Poisson distribution are equal, we say that the Poisson distribution satisfies the equi-dispersion property. This property is often violated in real-life count data. We have over-dispersion (or under-dispersion) when the variance is greater (or less) than the mean.

When the principle of complete randomness fails (that is the data is either over or under-dispersed), it is wise to use such a probabilistic model which can handle such curious situations. The basic alternative models for Poisson are Negative Binomial (NB) (Johnson, Kemp, and Kotz 2005) and generalized Poisson (GP) model (Consul and Famoye 1989). The main attribute of these two distributions as compared to Poisson distribution is that they have an additional parameter, sometimes called dispersion parameter, which makes them flexible. In the case of NB distribution , the additional parameter introduces over-dispersion and in the GP model , over- or under-dispersion character is incorporated by the additional parameter.

In addition to these alternative two distributions, there is vast literature available in order to deal with over-dispersion including some new (alternative) two- and three-parameter discrete distributions. Recently Déniz et al. (Gómez-Déniz, Vázquez-Polo, and Garcia 2017) introduced a simple count distribution (namely Modified Borel- Tanner (MBT) distribution) which has some eye catching properties like: (I) the distribution consists of one parameter; (II) it is one of the member of exponential family of distributions; (III) it belongs to the class of power series distribution; (IV) it is infinitely divisible; (V) it is unimodal; and (VI) variance larger than the mean, indicating that the one-parameter MBT distribution may be useful to model over-dispersed data. For full description about the MBT model, readers may please refer to Déniz et al. (Gómez-Déniz *et al.* 2017).

A discrete random variable X has a MBT distribution, if its probability mass function (PMF) is given by

$$P(X=x) = \frac{\Gamma(2x+1)}{\Gamma(x+2)\Gamma(x+1)} \frac{\alpha^x}{(1+\alpha)^{2x+1}}, \quad x = 0, 1, \cdots, \tag{1}$$

where $\alpha \in (0,1)$, and the notation used is $X \sim \mathbf{MBT}(\alpha)$. This distribution can also be re-written as

$$P(X=x) = C_x \frac{\alpha^x}{(1+\alpha)^{2x+1}}, \quad x = 0, 1, \cdots, \tag{2}$$

where

$$C_x = \frac{1}{x+1}\binom{2x}{x} \tag{3}$$

are the Catalan numbers denoted by C_n (Olver, Lozier, Boisvert, and Clark 2010) and the first few Catalan numbers are as: $C_0 = C_1 = 1$, $C_2 = 2$, $C_3 = 5$, $C_4 = 14$, $C_5 = 42$, $C_6 = 132$, $C_7 = 429$, $C_8 = 1430$, $C_9 = 4862$ and $C_{10} = 16796$.

From (1),the mean and variance are given by

$$E(X) = \frac{\alpha}{1-\alpha}, \tag{4}$$

$$Var(X) = \frac{\alpha(1+\alpha)}{(1-\alpha)^3}, \tag{5}$$

In a regression model framework, it is typically more useful to model the mean of the response variable. So, to obtain a regression structure for the mean of the MBT distribution, we shall work with a different parameterization of the MBT mass probability function. Let $\mu = \frac{\alpha}{(1-\alpha)}$ and hence $\alpha = \frac{\mu}{1+\mu}$. Then it follows from (4) and (5) that

$$\mathbb{E}(Y) = \mu \quad \& \quad \mathbb{V}(Y) = \mu(1+\mu)(1+2\mu),$$

where $\mu > 0$ is the mean of the response variable Y. The new re-parametrized mass function of $MBT(\mu)$ distribution is written as:

$$Pr(Y=y) = C_y \mu^y \frac{(1+\mu)^{y+1}}{(1+2\mu)^{2y+1}}, \quad y = 0, 1, \cdots, \tag{6}$$

where $\mu > 0$ and C_y are already defined in (3). Moreover,

$$\frac{\Pr(Y=y)}{\Pr(Y=y-1)} - \frac{\Pr(Y=y+1)}{\Pr(Y=y)} = \frac{-6(1+\mu)}{(1+2\mu)^2} \frac{1}{(y+1)(y+2)} < 0,$$

we have that the distribution is log-convex (infinitely divisible) and has decreasing failure rate (DFR). The fact that $\frac{\Pr(Y=y)}{\Pr(Y=y-1)}$, $y = 1, 2, ...$, forms a monotone increasing sequence requires that $Pr(Y=y)$ be a decreasing sequence in y. Therefore, the distribution is unimodal with

modal value on zero. In addition to it, the index of dispersion ($\mathbb{ID}$) which is actually ratio of variance to mean is

$$\mathbb{ID} = 1 + 3\mu + 2\mu^2 > 1,$$

It is interesting to note that no additional parameter in the MBT model is necessary to deal with over-dispersion, which makes the MBT model more parsimonious than other two- and three-parameters distributions used to model data with over-dispersion.

It is already well known that sometimes count data posses extra proportion of zeros. One can find ample amount of instances where count data exhibiting zero inflation is seen in various fields like medical science, public health, environmental sciences, agriculture and manufacturing applications. Zero-inflation, an indication of over-dispersion most frequently, means that the incidence of more zero counts than expected. A simple histogram or frequency plot with a large spike at zero gives an early warning of possible zero inflation. The basic and standard model for zero inflation is Zero-inflated Poisson (ZIP) model. The basic theory behind the derivation of the ZIP model is to mix a distribution degenerate at zero with a Poisson distribution. Since one could theoretically mix the degenerate distribution with any count distribution, we refer to the latter (non-degenerate) distribution/model as the baseline model. Also, over-dispersion can be the result of excess zeros or some other cause. In any case, the result is excess variability. In some cases, the ZIP model may not be appropriate for such data, since the baseline (Poisson) model does not accommodate the remaining over-dispersion not accounted for through zero-inflation. Additionally, it has been established that the ZIP parameter estimates can be severely biased if the nonzero counts are over-dispersed in relation to the Poisson distribution, leading to serious underestimation of standard errors and misleading inference for the regression parameters.

The motivation behind proposing the zero-inflated version of MBT model is that; it consists of only one parameter, the probabilities of the model are monotonically decreasing in x, it is an over-dispersed model and it has very simple closed form expressions which are easy to deal with. Mixing a distribution degenerate at zero with a baseline MBT distribution, we will propose the zero-Inflated Modified Borel–Tanner (**ZIMBT**) model given by

$$Pr(Y = y) = \begin{cases} \pi + (1-\pi)\frac{(1+\mu)}{(1+2\mu)}, & y = 0, \\ (1-\pi)C_y\mu^y\frac{(1+\mu)^{y+1}}{(1+2\mu)^{2y+1}}, & y > 0, \end{cases} \tag{7}$$

where C_y is already defined above in (3), $\mu > 0$ and $\pi \in (0, 1)$. If Y follows **ZIMBT** distribution with parameters μ and π, then the notion used is $Y \sim \mathbf{ZIMBT}(\mu, \pi)$. It can be noted that the PMF of **ZIMBT** given in 7 is very easy to handle, as it does not involve any complicated function at all. Moreover, the mean and variance of $\mathbf{ZIMBT}(\mu, \pi)$ are obtained as:

$$\mathbb{E}(Y) = (1-\pi)\mu, \tag{8}$$

$$\mathbb{V}(Y) = (1-\pi)\mu\left[\mu(1+\mu)(1+2\mu) + \pi\mu\right] \tag{9}$$

Since

$$\mathbb{ID} = \mu\left[(1+\mu)(1+2\mu) + \pi\right], \quad \mu > 0, \quad \pi \in (0, 1).$$

Note that the proposed distribution is zero-inflated. To confirm this it can be observed that zero Inflation Index (Puig and Valero 2006) is $Z_i = 1 + \frac{1}{(1-\pi)\mu}\log\left(\frac{\mu(1+\pi)+1}{1+2\mu}\right) > 0$.
Additionally, the probability generating function (pgf) of $Y \sim \mathbf{ZIMBT}(\mu, \pi)$ is

$$P_Y(t) = \pi + (1-\pi)\left[\frac{2(1+\mu)}{\mu + \sqrt{1 - 4(t-1)\mu(1+\mu)} + 3}\right], \quad |t| < 0. \tag{10}$$

In this paper, we also propose a new zero-inflated regression model on the basis of the **ZIMBT** distribution. So, having accounted for zero-inflation, if the data continue to suggest additional

over-dispersion, one may consider the **ZIMBT** model instead of the ZIP model. Similar to the ZIP and ZINB regression model setup, the parameters μ and π are related to covariates (explanatory variables). Furthermore, some quantities (e.g., score function, Fisher information matrix, etc.) related to the **ZIMBT** regression are simple and compact, which makes the frequenistic approach very easy to implement.

The rest of the paper is structured are as follows: In section 2 we present the review of data and its preliminary analysis including Poisson and ZIP regression model. The zero-inflated Modified Borel-Tanner regression model is developed in section 3 and estimation of parameters via ML method are being done in 3.1. An application of the proposed model on a real data is shown in section 4 followed by concluding remarks in the last section 5.

2. Review of data and preliminary analysis

The data is taken from Crawley (2012). This data has also been used by Lemonte et al. (Lemonte, Moreno-Arenas, and Castellares 2019). The data consists of count of infected blood cells per square millimetre on microscope slides prepared from randomly selected individuals of size 511. The explanatory variables are smoker (logical: yes or no), age (three levels: under 20, 21 to 59, 60 and over), sex (male or female) and body mass score (three levels: normal, overweight, obese). It is worth to mention that most of the patients, 314 individuals (approximately 61.4%) showed no damaged cells, and the maximum of 7 infected cells was observed in just two patients (#314 & #246). It is also evident from the preliminary view of data that smokers had a substantially higher mean count than non-smokers. Initially we will consider Poisson regression model as:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{smoking}_i + \beta_2 \text{gender}_i + \beta_3 \text{age}_i + \beta_4 \text{weight}_i, \quad i = 1, 2, \ldots, 511.$$

From Table 1, it is evident that regressors like *gender* is marginally non-significant while

Table 1: Parameter estimates: Poisson regression

Parameter	Estimate	t-value	p-value
β_0	-0.2200	-0.802	0.4225
β_1	-1.1916	-11.399	0.0000
β_2	0.2016	1.950	0.0512
β_3	0.0075	0.118	0.9061
β_4	0.2626	4.679	0.0000

as *age* is highly non-significant, therefore we will revise our model by including only three explanatory variables; smoker, gender and weight. Now, the revised Poisson regression model is:

$$\log(\mu_i) = \beta_0 + \beta_1 \text{smoking}_i + \beta_2 \text{gender}_i + \beta_3 \text{weight}_i, \quad i = 1, 2, \ldots, 511.$$

The statistical quantities like log- likelihood (LL), AIC (Akaike Information Criterion) (Akaike 1974) and BIC (Bayesian Information Criterion) (Schwarz *et al.* 1978) criteria are presented in Table 2. In addition to it, the parameter estimates, their standard error (SE), and asymptotic 95% confidence interval (CI) of the Poissom regression model are given by Table 2. Usually, regression coefficients represents the mean change in response variable for one unit change in the predictor variable while holding the other predictors in the model constant. Therefore, from Table 2, the coefficient for *smoking* indicates that for every additional smoker, we can expect count of infected blood cells to decrease by an average of 1.1947. Other regression coefficients of the model can be interpreted similarly as well. Further corresponding to the estimates, their respective t-value and p-value are given in Table 3. The residual deviance equals 851.85 , which is much greater than the residual degrees of freedom (i.e. 507), indicating over-dispersion and hence indicates that the Poisson regression model is not suitable to model the data.

Table 2: Parameter estimates: Poisson regression

Parameter	Estimate	SE	95% CI
β_0	-0.2016	0.2250	(-0.642 ; 0.239)
β_1	-1.1947	0.1011	(-1.393;-0.997)
β_2	0.2000	0.1024	(-0.001 ;0.401)
β_3	0.2630	0.0560	(0.153;0.373)
LL	-682.7633		
AIC	1373.527		
BIC	1390.472		

Table 3: Parameter estimates: Poisson regression

Parameter	Estimate	t-value	p-value
β_0	-0.2016	-0.896	0.3700
β_1	-1.1947	-11.818	0.0000
β_2	0.2000	1.952	0.0510
β_3	0.2630	4.697	0.0000

As already mentioned earlier, classical Poisson model is the first choice to model count data, but because of lack of fit and over-dispersion which may be due to presence of large number of zeros in the sample, it is not always a suitable choice. So, in order to deal with extra proportion of zeros and over-dispersion, we shall move towards ZIP regression model given by

$$\log(\mu_i) = \beta_0 + \beta_1 \text{smoking}_i + \beta_2 \text{gender}_i + \beta_3 \text{weight}_i,$$
$$\log\left(\frac{\pi_i}{1+\pi_i}\right) = \gamma_0 + \gamma_1 \text{smoking}_i + \gamma_2 \text{gender}_i + \beta_3 \text{weight}_i,$$

with $i = 1, 2, \ldots, 511$. Table 4 shows the ML estimates of the parameters, LL, AIC, BIC values and 95% asymptotic CI's. Also the t-values and p-values of the estimates are being shown in Table 5. On the basis of LL, AIC and BIC values, it clearly indicates that ZIP model outperforms Poisson regression model for the said data. In the next Section, we will introduce the use of the **ZIMBT** regression model introduced in this paper to model these data improves considerably the fit in terms of model fitting.

Table 4: Parameter estimates: ZIP regression & 95% CI

Parameter	Estimate	SE	95% CI
β_0	0.3939	0.2893	(-0.1731; 0.9609)
β_1	-0.4852	0.1228	(-0.7259 ; -0.2445)
β_2	0.0274	0.1235	(-0.2148; 0.2695)
β_3	0.2471	0.0715	(0.1068; 0.3873)
γ_0	-0.4320	0.6109	(-1.6295; 0.7655)
γ_1	1.6308	0.3055	(1.0319; 2.2296)
γ_2	-0.3932	0.2574	(-0.8977; 0.1112)
γ_3	-0.0369	0.1544	(-0.3394;0.2656)
LL	-605.253		
AIC	1226.505		
BIC	1232.28		

3. ZIMBT regression model

Let $Y_1, Y_2, ..., Y_n$ be n independent random variables, where each Y_i, for $i = 1, 2, ..., n$, follows the PMF (7) with mean μ_i and probability π_i; that is, $Y_i \sim \mathbf{ZIMBT}(\mu_\mathbf{i}, \pi_\mathbf{i})$, (for $i = 1, 2, ..., n$).

Table 5: Parameter estimates: ZIP regression

Parameter	Estimate	t-value	p-value
β_0	0.3939	1.362	0.1733
β_1	-0.4852	-3.951	0.0001
β_2	0.0274	0.222	0.8246
β_3	0.2471	3.453	0.0006
γ_0	-0.4320	-0.719	0.480
γ_1	1.6308	5.338	0.0000
γ_2	-0.3932	-1.528	0.1270
γ_3	-0.0369	-0.239	0.8110

Suppose the following functional relation is established here:

$$\log(\mu_i) = \nu_{1i} = \mathbf{x}_i^\top \beta, \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \nu_{2i} = \mathbf{s}_i^\top \gamma, \tag{11}$$

with $\beta = (\beta_1, \beta_2, ..., \beta_p)^\top$ and $\gamma = (\gamma_1, \gamma_2, ..., \gamma_q)^\top$ are vectors of unknown regression coefficients which are supposed to be functionally independent, $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}^q$ with $p+q < n$, ν_{1i} and ν_{2i} are linear predictors, and $\mathbf{x}_i^\top = (x_{i1}, x_{i2}, ..., x_{ip})$ and $\mathbf{s}_i^\top = (s_{i1}, s_{i2}, ..., s_{iq})$ are values of the observations on p and q known covariates (or independent variables or regressors). Also, the matrices $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]^\top$ and $\mathbf{S} = [\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_n}]^\top$ have rank p and q, respectively. Moreover, generally in practice, we have $x_{i1} = s_{i1} = 1$ (for $i = 1, 2, ..., n$), which corresponds to the intercept. Commonly covariates in $\mathbf{S}$ are subset of the covariates in $\mathbf{X}$ (but not necessary). It is pertinent to mention here that similar results could be obtained for other link functions in (11), but, the log and logit link functions are most common in such a case.

3.1. Parameter estimation

The ML method is considered to estimate the parameter vector $\Theta = (\beta^\top, \gamma^\top)^\top$. Let $\mathbf{y} = (y_1, y_2, ..., y_n)^\top$ be the vector of the observed responses. The log-likelihood function, except for the constant terms, is given by

$$\begin{aligned} l(\Theta) &= \sum_{y_i:y_i=0} \log\left[e^{\nu_{2i}} + \frac{(1+\mu_i)}{(1+2\mu_i)}\right] + \sum_{y_i:y_i>0} y_i \log(\mu_i) + \sum_{y_i:y_i>0} (y_i+1)\log(1+\mu_i) \\ &\quad + \sum_{y_i:y_i>0} (2y_i+1)\log(1+2\mu_i) - \sum_{i=1}^{n} \log\left(1+e^{\nu_{2i}}\right), \end{aligned} \tag{12}$$

with $\mu_i = e^{\nu_{2i}} = e^{{X_i}^\top \beta}$ for $i = 1, 2. \ldots, n$. The ML estimator $\hat{\Theta} = (\hat{\beta}^\top, \hat{\gamma}^\top)^\top$ of $\Theta = (\beta^\top, \gamma^\top)^\top$, where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^\top$ & $\hat{\gamma} = (\hat{\gamma_1}, \hat{\gamma_2}, \ldots, \hat{\gamma_q})^\top$ can be find by maximising log-likelihood function $l(\Theta)$ (12) with respect to parameter vector $\Theta = (\beta^\top, \gamma^\top)^\top$. Under some mild regularity conditions, as $n \to \infty$ (n is sample size), the ML estimator Θ of $\hat{\Theta}$ is unique and asymptotically normal (Cox and Hinkley 1974), which will be discussed in detail a bit later in this section.

The Score function, denoted by $S(\beta, \gamma)$, is obtained by taking first derivative of $l(\Theta)$ with respect to some unknown parameters, here $S(\beta, \gamma)$ is $(p+q)$ vector such that $S(\beta, \gamma) = (S_\beta(\beta, \gamma)^\top, S_\gamma(\beta, \gamma)^\top)^\top$, where $S_\beta(\beta, \gamma) = X^\top \zeta$, $S_\gamma(\beta, \gamma) = S^\top \Lambda$, $\zeta = (\zeta_1, \zeta_2, \ldots, \zeta_n)^\top$ and $\Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_n)^\top$, with

$$\begin{aligned} \zeta_i &= \begin{cases} -\frac{e^{\nu_{1i}}}{(1+2e^{\nu_{1i}})[1+e^{\nu_{2i}}+e^{\nu_{1i}}+2e^{\nu_{2i}}e^{\nu_{1i}}]}, & y_i = 0, \\ y_i + \frac{(y_i+1)e^{\nu_{1i}}}{(1+e^{\nu_{1i}})} - \frac{2(2y_i+1)e^{\nu_{1i}}}{(1+2e^{\nu_{1i}})}, & y_i > 0, \end{cases} \\ \Lambda_i &= \frac{e^{\nu_{2i}} I_{(y_i=0)}}{\left[e^{\nu_{2i}} + \frac{(1+\mu_i)}{(1+2\mu_i)}\right]} - \frac{e^{\nu_{2i}}}{(1+e^{\nu_{2i}})}, \end{aligned} \tag{13}$$

where $I(\cdot)$ denotes the indicator function. The ML estimates $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^\top$ and $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_q)^\top$, of $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)^\top$ respectively, can also be obtained by solving simultaneously the nonlinear system of equations $S_\beta(\beta, \gamma) = 0_p$ and $S_\gamma(\beta, \gamma) = 0_q$, where 0_k denotes a k-dimensional vector of zeros. These non-linear system of equations are not in closed forms and hence can not be solved directly, therefore we have to make use of some iterative algorithm , like Newton Raphson method, to get the ML estimates. While using any iterative algorithm, it is always of immediate interest to choose the appropriate initial values. For obtaining the the parameter estimates of the proposed **ZIMBT** regression model we took the initial values from the estimates obtained from the ZIP regression model, which can be easily obtained through the R function `zeroinfl(···)` in the `pscl` library [Jackman (2015), Zeileis, Kleiber, and Jackman (2008)].

Since the new parametric **ZIMBT** regression model corresponds to a regular ML problem, regular in the sense that it satisfies all the regularity conditions which are as: (i) the pmf is distinct (ii) the pmfs have common support for all θ (iii) The point θ_0, is the real parameter that is, is an interior point in some set (Ω). These three conditions together guarantees that the likelihood is maximised at the true parameter θ_0 and then that the mle $\hat{\theta}$ that solves the $\frac{\partial l}{\partial \theta} = 0$ is conistent. (iv) The $p(x|\theta)$ is twice differentiable as a function of θ.

We have that the standard asymptotics apply; that is, the ML estimators of the model parameters are asymptotically normal, asymptotically unbiased and have asymptotic variance-covariance matrix given by the inverse of the expected Fisher information matrix. Let $K(\beta, \gamma)$ be the $(p+q) \times (p+q)$ expected Fisher information matrix for $K(\beta, \gamma)$. Thus, we have

$$\hat{\Theta} \overset{a}{\sim} \mathbb{N}_{p+q}\left(\Theta, K(\beta, \gamma)^{-1}\right),$$

where $\overset{a}{\sim}$ means approximately distributed. It is important to find the mathematical expression for $K(\beta, \gamma)$ which can be used to obtain asymptotic Standard Errors (SEs) for the ML estimates. After some calculation, the expected Fisher information matrix for (β, γ) takes the form

$$K(\beta, \gamma) = \begin{bmatrix} X^\top D_1 X & X^\top D_2 S \\ S^\top D_2 X & S^\top D_3 S \end{bmatrix},$$

where $\mathbf{D_1}$=diag$\{d_{1i}\}$, $\mathbf{D_2}$=diag$\{d_{2i}\}$ and $\mathbf{D_3}$=diag$\{d_{3i}\}$ stands for a diagonal matrix with typical element $b_i (i = 1, 2, \ldots . n)$. All quantities necessary to compute the above matrices are given below:

$$d_{1i} = \begin{cases} d_{1i}^{(0)}, & y_i = 0, \\ d_{1i}^{(c)}, & y_i > 0, \end{cases}$$

where

$$\begin{aligned}
d_{1i}^{(0)} &= -\frac{e^{\nu_{1i}}\left(4e^{\nu_{2i}}e^{\nu_{1i}} + 2e^{2\nu_{1i}} - e^{\nu_{2i}} - 1\right)}{(1+2e^{\nu_{2i}})^2\left[1 + e^{\nu_{2i}} + e^{\nu_{1i}} + 2e^{\nu_{2i}}e^{\nu_{1i}}\right]^2}, \\
d_{1i}^{(c)} &= -\frac{(y_i+1)e^{\nu_{1i}}}{(1+e^{\nu_{1i}})^2} + \frac{2(2y_i+1)e^{\nu_{1i}}}{(1+2e^{\nu_{1i}})^2}, \\
d_{2i} &= \begin{cases} -\frac{e^{\nu_{1i}+\nu_{2i}}}{\left[1+e^{\nu_{1i}}+e^{\nu_{2i}}+2e^{\nu_{1i}+\nu_{2i}}\right]^2}, & y_i = 0, \\ 0, & y_i > 0, \end{cases} \\
d_{3i} &= -\frac{e^{\nu_{2i}}(1+\mu_i)(1+2\mu_i)I_{(y_i=0)}}{\left[1+\mu_i+e^{\nu_{2i}}(1+2\mu_i)\right]^2} + \frac{e^{\nu_{2i}}}{(1+e^{\nu_{2i}})^2}.
\end{aligned}$$

The above asymptotic normal distribution can be used to construct approximate Confidence Intervals (CIs) for the parameters. Let $\beta_j (j = 1, 2, \ldots, p)$ and $\gamma_k (k = 1, 2, \ldots, q)$ be the j-th and k-th components of β and γ, respectively. For $0 < \alpha < \frac{1}{2}$, the asymptotic CIs $\hat{\beta}_j \pm z_{(1-\alpha/2)} S.E(\hat{\beta}_j)$ and $\hat{\gamma}_k \pm z_{(1-\alpha/2)} S.E(\hat{\gamma}_k)$ for β_j and γ_k , respectively, both with asymptotic coverage of $100(1-\alpha)\%$. Here, $S.E(\cdot)$ is the square root of the diagonal element of $K(\hat{\beta}, \hat{\gamma})^{-1}$

corresponding to each parameter (i.e. the asymptotic S.E), and $z_{(1-\alpha/2)}$ denotes the $(1-\alpha/2)$-th quantile of the standard normal distribution.

4. Numerical illustration

In this section, we will examine the application of **ZIMBT** regression model on the real data set whose preliminary information is already discussed in section 2. Throughout this article, all the computations were carried with the help of R Software. Before going to **ZIMBT** model, let us start from $Y_i \sim MBT(\mu_i)$ distribution whose PMF is already given by (6).

$$\log(\mu_i) = \beta_0 + \beta_1 \text{smoking}_i + \beta_2 \text{gender}_i + \beta_3 \text{weight}_i, \quad i = 1, 2, \ldots, 511.$$

Table 6 lists the ML estimates, asymptotic SEs, and the 95% asymptotic CIs and Table 7 lists ML estimates, t-values and p-values for the MBT regression parameters. Note that the MBT regression outperforms the Poisson regression on the basis of the maximum likelihood value, AIC and BIC values. Also it is evident from Table 7 gender is statistically non-significant.

Table 6: Parameter estimates: MBT regression & 95% CI

Parameter	Estimate	SE	95% CI
β_0	0.0685	0.3091	(-0.537 ;0.674)
β_1	-1.1215	0.2672	(-1.645 ;-0.598)
β_2	0.2204	0.2344	(-0.239 ; 0.680)
β_3	0.6566	0.2633	(0.141;1.173 8)
LL	-657.46		
AIC	1334.9		
BIC	1377.27		

Table 7: Parameter estimates: MBT regression model

Parameter	Estimate	t-value	p-value
β_0	0.0685	0.2217	0.8246
β_1	-1.1215	-4.1965	0.000
β_2	0.2204	0.9402	0.3471
β_3	0.6566	2.4936	0.0126

Next, we will take now $Y_i \sim ZIMBT(\mu_i, \pi)$ as:

$$\begin{aligned} \log(\mu_i) &= \beta_0 + \beta_1 \text{smoking}_i + \beta_2 \text{gender}_i + \beta_3 \text{weight}_i, \\ \log\left(\frac{\pi_i}{1+\pi_i}\right) &= \gamma_0 + \gamma_1 \text{smoking}_i + \gamma_2 \text{gender}_i + \gamma_3 \text{weight}_i, \end{aligned}$$

Table 8 shows the ML estimates, asymptotic SEs, and the 95% asymptotic CIs. Also the corresponding t-values and p-values with respect to their estimates are presented by Table 9. From Table 8 , one can clearly claim that the **ZIMBT** regression model outperforms the Poisson, ZIP and MBT regression model in terms of the maximum likelihood value, AIC and BIC values. Also it is evident from Table 9 that the gender is statistically non-significant in the zero and count components.

Since the familiar and immediate choice for count data regression analysis after the Poisson and ZIP regression models is NB and Zero-inflated Negative Binomial (ZINB) regression model, so we will fit the data by using both NB and ZINB model. The PMF of NB model is given by

$$P(Y = y) = \left(\frac{\phi}{\phi+\mu}\right)^{\phi} \left(\frac{\mu}{\phi+\mu}\right)^{y} \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}, \quad y = 0, 1, \ldots, \tag{14}$$

Table 8: Parameter estimates: **ZIMBT** regression & 95% CI

Parameter	Estimate	SE	95% CI
β_0	0.0729	0.2695	(-0.455 ; 0.601)
β_1	-1.1104	0.0876	(-1.282; -0.939)
β_2	0.1898	0.2632	(-0.326 ; 0.706)
β_3	0.7038	0.5333	(-0.341 ;1.749)
γ_0	-22.6387	797.7407	(-1586.210;1540.933)
γ_1	12.3109	791.1154	(-1538.275;1562.897)
γ_2	-14.3000	2012.7340	(-3959.259;3930.659)
γ_3	7.8987	102.6810	(-193.356;209.153)
LL	-460.37		
AIC	940.75		
BIC	983.11		

Table 9: Parameter estimates: **ZIMBT** regression model

Parameter	Estimate	t-value	p-value
β_0	0.0729	0.2704	0.7868
β_1	-1.1104	-12.6782	0.0000
β_2	0.1898	0.7213	0.4707
β_3	0.7038	1.3197	0.1869
γ_0	-22.6387	-0.0284	0.9774
γ_1	12.3109	0.0156	0.9876
γ_2	-14.3000	-0.0071	0.9943
γ_3	7.8987	0.0769	0.9387

where $\mu > 0, \phi > 0$ and $\Gamma(\cdot)$ is the gamma function such that $\Gamma(n) = (n-1)!$. Note that ϕ is called dispersion parameter and as $\phi \to \infty$, NB reduces to Poisson distribution. The mean and variance of (14) is μ and $(\mu + \frac{\mu^2}{\phi})$. Similarly the PMF of ZINB is given by

$$Pr(Y=y) = \begin{cases} \pi + (1-\pi)\left(\frac{\phi}{\phi+\mu}\right)^{\phi}, & y = 0, \\ (1-\pi)\left(\frac{\phi}{\phi+\mu}\right)^{\phi}\left(\frac{\mu}{\phi+\mu}\right)^{y}\frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}, & y > 0, \end{cases} \tag{15}$$

where $\mu > 0, \phi > 0$ and $0 < \pi < 1$. Reader may please refer to Hilbe (Hilbe 2011) for further details. we use the notations as $Y_i \sim NB(\mu_i, \phi)$ and $Y_i \sim ZINB(\mu_i, \phi, \pi_i)$, i.e.,

$$\log(\mu_i) = \beta_0 + \beta_1 \text{smoking}_i + \beta_2 \text{gender}_i + \beta_3 \text{weight}_i,$$
$$\log\left(\frac{\pi_i}{1+\pi_i}\right) = \gamma_0 + \gamma_1 \text{smoking}_i + \gamma_2 \text{gender}_i + \gamma_3 \text{weight}_i,$$

with $i = 1, 2, \ldots, 511$. Regarding the NB regression, one can make use of R function `glm.nb`$(\cdots)$ from the MASS library (Venables and Ripley 2002), where as the R function `zeroinfl`$(\cdots)$ in the `pscl` library can be used for the ZINB regression. The estimates corresponding to the NB and ZINB regression models are listed in Tables 10 and 11, respectively. From Tables 10 and 11, note that the ZINB regression model provides an enhancement over the NB regression model on the basis of the maximum likelihood value, AIC and BIC.

The results of all the competing models along with **ZIMBT** regression model taken in to consideration in this article are presented in a single Table 12, which shows that the **ZIMBT** model outbeats Poisson, ZIP, NB, ZINB and MBT models and provides a good fit in comparison. Therefore, it is clearly evident from the data analysis that the proposed **ZIMBT** regression model should be preferred.

Table 10: Parameter estimates: NB regression model

Parameter	Estimate (S.E)	95% CI	t-value	p-value
β_0	-0.3039 (0.3382)	(-0.967;0.359)	-0.8985	0.3689
β_1	-1.1574 (0.1587)	(-1.468 ;-0.846)	-7.2936	0.0000
β_2	0.2578 (0.1540)	(-0.044;0.560)	1.6745	0.0940
β_3	0.2586 (0.0876)	(0.087;0.430)	2.9518	0.0032
LL	-624.3465			
AIC	1258.693			
BIC	1279.875			

Table 11: Parameter estimates: ZINB regression model

Parameter	Estimate (S.E)	95% CI	t-value	p-value
β_0	0.2305 (0.3584)	(0.626; 1.122)	3.102	0.0019
β_1	-0.4951 (0.1409)	(-0.767;-0.283)	-3.557	0.0004
β_2	0.0531 (0.1435)	(-0.243; 0.243)	0.003	0.9979
β_3	0.2789 (0.0887)	(-1.735;-0.371)	-1.000	0.3170
γ_0	-0.8031 (0.7867)	(1.058; 2.286)	4.463	0.0000
γ_1	1.7683(0.3935)	(-0.923; 0.029)	-1.541	0.123
γ_2	-0.3909(0.2868)	(-0.923; 0.029)	-1.541	0.123
γ_3	0.0179(0.1817)	(-0.923; 0.029)	-1.541	0.123
$\log(\phi)$	1.97795(0.7278)			
LL	-603.6473			
AIC	1225.295			
BIC	1263.422			

Table 12: Summaries of fitting measures results for the models considered

Criterion	Poisson	ZIP	NB	ZINB	MBT	ZIMBT
LL	-682.763	-605.253	-624.3465	-603.6473	-657.46	-460.37
AIC	1373.527	1226.505	1258.693	1225.295	1334.9	940.75
BIC	1390.472	1232.28	1279.875	1263.422	1377.27	983.11

5. Conclusions

With the introduction of zero-inflated Poisson regression model by Lambert (1992), there is a significant growing interest, both in the econometrics and statistics literature, in zero-inflated models. In short, zero-inflated models are mixture models that combine a count component and a point mass at zero. Thus, there are two sources of zeros: zeros may come from both the point mass and from the count component. This paper has introduced the zero-inflated version of the already existing MBT model, which is actually a modified version of Borel-Tanner distribution. ML technique have been employed for the estimation of parameters Finally, we illustrate the methodology developed in this paper by means of an application to real data. The **ZIMBT** regression model seems to be an interesting model in practice when compared with some familiar regression models (including both zero-inflated or not).

References

Akaike H (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Consul PC, Famoye F (1989). "The Truncated Generalized Poisson Distribution and Its Estimation." *Communications in Statistics-Theory and Methods*, **18**(10), 3635–3648.

Cox DR, Hinkley DV (1974). *Theoretical Statistics*. Chapman and Hall, London.

Crawley MJ (2012). *The R Book*. John Wiley and Sons, New York.

Gómez-Déniz E, Vázquez-Polo FJ, Garcia V (2017). "The Modified Borel-Tanner (MBT) Regression Model." *REVSTAT Statistical Journal*.

Hilbe JM (2011). *Negative Binomial Regression*. Cambridge University Press.

Jackman S (2015). "pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory (R package version 1.4. 9). Department of Political Science." URL http://pscl.stanford.edu.

Johnson NL, Kemp AW, Kotz S (2005). *Univariate Discrete Distributions*. John Wiley & Sons.

Lambert D (1992). "Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics*, **34**(1), 1–14.

Lemonte AJ, Moreno-Arenas G, Castellares F (2019). "Zero-inflated Bell Regression Models for Count Data." *Journal of Applied Statistics*.

Olver FWJ, Lozier DW, Boisvert RF, Clark CW (2010). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge university press.

Puig P, Valero J (2006). "Count Data Distributions: Some Characterizations with Applications." *Journal of the American Statistical Association*, **101**(473), 332–340.

Schwarz G, *et al.* (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, **6**(2), 461–464.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S (4th edition)*. Springer, New York.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *Journal of Statistical Software*, **27**(8), 1–25.

Affiliation:

Peer Bilal Ahmad
Department of Mathematical Sciences
Islamic University of Science and Technology
192122 Awantipora, J&k, India
E-mail: bilalahmadpz@gmail.com
URL: https://www.iust.ac.in/Index/EmployeeDetails.aspx?DeptCode=DOM&EmpId=1345

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 51
January 2022

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2020-08-31
Accepted: 2021-03-30