International Journal of Statistics and Applications 2019, 9(5): 134-142
DOI: 10.5923/j.statistics.20190905.02

A Combined Estimation Method to Estimate the Parameters of the Zero-One Inflated Negative Binomial Distributions

Rafid S. A. Alshkaki

Department of General Requirement, Ahmed Bin Mohammed Military College, Doha, Qatar

Abstract In this paper, the zero and one inflated negative binomial distributions is considered. A combined method of relative frequencies and maximum likelihood estimators was introduced to estimate the parameters of the zero and one inflated negative binomial distribution. A simulation study was conducted to check the performance of this estimation method using the mean squares error of each of the parameter estimates for six simulated different zero and one inflated negative binomial distributions models. The proposed estimation procedures was used to estimate the parameters of six real life data sets models and it gave good results.

Keywords Zero One Inflated Negative Binomial Distribution, Maximum likelihood Estimation, Relative Frequency, Non Negative Integer Sampling

1. Introduction

In recent researcher's literature statistical modelling work, frequencies of zeros may be significantly higher than the predicated frequency by the standard statistical models. This might be lead to wrong conclusions about the actual statistical model. Such models are called zero inflated models. Furthermore, frequencies of zeros and ones may be also jointly significantly higher than the predicated frequency by the standard statistical models also, leading to zeros and ones inflated models. Hence the problem of estimating the model's parameters may be need further work more than the classical statistical methods.

Gan (2000) studied the properties of the maximum likelihood estimates (MLE) of zero inflated model parameters, including their existence, uniqueness, strong consistency and asymptotic normality under regularity conditions. Preisser et al. (2012) considered reviews of the zero inflated Poisson and the zero inflated negative binomial (ZINB) regression models applied to dental caries, with emphasis on the description of the models and the interpretation of fitted model results given the study goals. Staub and Winkelmann (2012) noted that zero-inflated Poisson and the ZINB maximum likelihood estimators are not robust to misspecification, and proposed Poisson quasi-likelihood estimators, as an alternative, as consistent estimators in the presence of excess zeros without having to specify the full distribution. Phang et al. (2013) reviewed some literature on the zero inflated models and provide a variety of examples from different disciplines in the applications of zero inflated models, as well as, discussed different model selection methods used in model comparison. Astuti and Mulyanto (2016) used the MLE method to estimate the parameter on ZINB regression model through maximizing the likelihood function using expectation maximization algorithm. Lukusa et al. (2017) considered the zero-inflated models as the most appropriate approach for dealing properly with this issue of excess zeros, reviewed studies the missing data problem and the zero-inflated feature in modeling zero-inflated data, and discussed their methodologies and results and some potential directions of the future research. Yang et al. (2017) evaluated the performance of several models under different conditions of zero -inflation and dispersion, and used results from simulated and real data and showed, when data have excessive zeros and over-dispersion, that the zero-altered or ZINB model were preferred over others, such as, ordinary least-squares regression with log-transformed outcome, and Poisson model. Yusuf et al. (2017) used the values of Vuong z-statistic, -2logLL, AIC and BIC selection criteria to select the best fitted zero inflated Poisson and ZINB regression models, and suggested that the ZINB regression as the best model for predicting number of falls in the presence of excess zeros and over-dispersion. Zamri and Zamzuri (2017) reviewed the zero inflated models literature,

* Corresponding author:
rafid@abmmc.edu.qa (Rafid S. A. Alshkaki)
Published online at http://journal.sapub.org/statistics

provided a recent development and summary on models for count data with extra zeros, and they found in the literature that the most popular zero inflated models used are zero inflated Poisson and ZINB.

The zeros and ones inflated models are not given such attention in the literature as the zero inflated models. In particular, Alshkaki (2017) gave an exact form of the ME of the parameters zero-one inflated negative binomial distribution (ZOINBD), and found numerically that this method is generally not an accurate method to estimate the parameters of the ZOINB models and may lead to misleading predication.

In this paper, the definition of the ZOINBD was introduced in Section 2, followed in Section 3 by introduced a combined method of relative frequencies and MLE. In Section 4, we conducted a simulation study to check the performance of the proposed estimation procedure using the mean squires error computed from different sample sizes for the estimated parameters on six simulated different ZOINBD models. Finally, in Section 5, we used the proposed estimation procedure to estimate the parameters and the frequencies of six different real life data sets.

2. The Negative Binomial Distributions and Its Zero-One Inflated Form

Let $k > 0$ *and* $\theta \in (0,1)$, then the discrete random variable (rv) Y having a probability mass function (PMF) given by;

$$P(Y = y) = \begin{cases} \binom{k+y-1}{y} \theta^y (1-\theta)^k, & y = 0, 1, 2, 3, \dots \\ , & \text{otherwise,} \end{cases} \quad (1)$$

is said to have a negative binomial distribution (NBD) with parameters k and θ . We will denote that by writing $Y \sim NBD(k, \theta)$. See Johnson et al. (2005), for other forms and parameterizations of the NBD.

Let $\alpha \in (0,1)$ be a proportion of zero added to the rv Y, and let $\beta \in (0,1)$ be an extra proportion added to the proportion of ones of the rv Y, such that $0 < \alpha + \beta < 1$, then, the rv X defined by, Alshkaki (2017);

$$P(X = x) = \begin{cases} \alpha + (1-\alpha-\beta)\,(1-\theta)^k\,, & x = 0 \\ \beta + k(1-\alpha-\beta)\,\theta\,(1-\theta)^k\,, & x = 1 \\ (1-\alpha-\beta)\binom{k+x-1}{x}\theta^x\,(1-\theta)^k\,, & x = 2, 3, 4, \dots \\ 0\,, & \text{otherwise,} \end{cases} \quad (2)$$

is said to have a ZOINBD with parameters $k, \theta, \alpha,$ and β, and we will denote that by writing $X \sim ZOINBD(k, \theta; \alpha, \beta)$.

Alshkaki (2017) noted that, if $\beta \to 0$, then (2) reduces to the form of the ZINBD. Similarly, the case with $\alpha \to 0$ and $\beta \to 0$, reduces to the standard case of the NBD.

Although, it does not fit the nature of the supposed model, Alshkaki (2017) noted that, the inflation parameters α and β may also take negative values providing that

$$\alpha \in \left(\max\left\{ -1, -(1-\beta)\frac{(1-\theta)^k}{1-(1-\theta)^k} \right\}, 0 \right)$$

and

$$\beta \in \left(\max\left\{ -1, -(1-\alpha)\frac{k\theta\,(1-\theta)^k}{1-k\theta\,(1-\theta)^k} \right\}, 0 \right)$$

without violating that (2) is a PMF. This situation represents the excluding proportion of zero's and one's, respectively, from the standard model given by (1).

3. Maximum Likelihood and Relative Frequencies Estimators

Let $x_1, x_2, \dots, x_n$ be a random sample from ZOINBD as given by (2), and let for i=1, 2, … n,

$$\alpha_i = \begin{cases} 1 & \text{if } x_i = 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\beta_i = \begin{cases} 1 & \text{if } x_i = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$c(k, x) = \binom{k+x-1}{x}$$

then, for i=1, 2, … , n, (2) can be written, for $x_i = 0, 1, 2, \dots$, in the following form;

$$P(X_i = x_i) = \{\alpha + (1-\alpha-\beta)(1-\theta)^k\}^{\alpha_i}\{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k\}^{\beta_i}\ \{(1-\alpha-\beta)c(k, x_i)\theta^{x_i}(1-\theta)^k\}^{1-\alpha_i-\beta_i}$$

Hence, the likelihood function $L = L(\theta, \alpha, \beta; x_1, x_2, \dots, x_n)$ can be written as;

$$L = \prod_{i=1}^{n}\{\alpha + (1-\alpha-\beta)(1-\theta)^k\}^{\alpha_i}\{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k\}^{\beta_i}\ \{(1-\alpha-\beta)c(k, x_i)\theta^{x_i}(1-\theta)^k\}^{1-\alpha_i-\beta_i}$$

$$= \{\alpha + (1-\alpha-\beta)(1-\theta)^k\}^{n_0}\{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k\}^{n_1}\prod_{i=1}^{n}\{(1-\alpha-\beta)c(k, x_i)\theta^{x_i}(1-\theta)^k\}^{c_i}$$

where $c_i = 1 - \alpha_i - \beta_i$, $n_0 = \sum_{i=1}^{n}\alpha_i$, and $n_1 = \sum_{i=1}^{n}\beta_i$. Note that n_0 and n_1 represent, respectively, the number of zeros and the number of ones in the sample. Therefore,

$$\log L = n_0 \log\{\alpha + (1-\alpha-\beta)(1-\theta)^k\} + n_1 \log\{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k\} + \sum_{i=1}^{n} c_i \log\left((1-\alpha-\beta)c(k,x_i)\theta^{x_i}(1-\theta)^k\right)$$

$$\log L = n_0 \log\{\alpha + (1-\alpha-\beta)(1-\theta)^k\} + n_1 \log\{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k\} + \sum_{i=1}^{n} c_i \log(1-\alpha-\beta) + \sum_{i=1}^{n} c_i \log[c(k,x_i)] + \sum_{i=1}^{n} c_i x_i \log(\theta) + k\sum_{i=1}^{n} c_i \log(1-\theta)$$

It follows that,

$$\frac{\partial}{\partial\alpha}\log L = \frac{n_0\left[1-(1-\theta)^k\right]}{\alpha + (1-\alpha-\beta)(1-\theta)^k} - \frac{n_1 k\theta(1-\theta)^k}{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k} - \frac{n_c}{(1-\alpha-\beta)} \tag{3}$$

where $n_c = \sum_{i=1}^{n} c_i \equiv n - n_0 - n_1$, and hence,

$$\frac{\partial^2}{\partial\alpha^2}\log L = -\frac{n_0\left[1-(1-\theta)^k\right]^2}{[\alpha + (1-\alpha-\beta)(1-\theta)^k]^2} - \frac{n_1\left[k\theta(1-\theta)^k\right]^2}{[\beta + (1-\alpha-\beta)k\theta(1-\theta)^k]^2} - \frac{n_c}{(1-\alpha-\beta)^2}$$

therefore, $\frac{\partial^2}{\partial\alpha^2}\log L < 0$, which indicates that L has a local maximum at α. Similarly,

$$\frac{\partial}{\partial\beta}\log L = -\frac{n_0(1-\theta)^k}{\alpha + (1-\alpha-\beta)(1-\theta)^k} + \frac{n_1\left[1-k\theta(1-\theta)^k\right]}{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k} - \frac{n_c}{1-\alpha-\beta}$$

and hence,

$$\frac{\partial^2}{\partial\beta^2}\log L = -\frac{n_0(1-\theta)^{2k}}{[\alpha + (1-\alpha-\beta)(1-\theta)^k]^2} - \frac{n_1\left[1-k\theta(1-\theta)^k\right]^2}{[\beta + (1-\alpha-\beta)k\theta(1-\theta)^k]^2} - \frac{n_c}{(1-\alpha-\beta)^2}$$

therefore, $\frac{\partial^2}{\partial\beta^2}\log L < 0$, which indicates that L has a local maximum at β. Next,

$$\frac{\partial}{\partial\theta}\log L = -\frac{n_0 k(1-\alpha-\beta)(1-\theta)^{k-1}}{\alpha + (1-\alpha-\beta)(1-\theta)^k} + \frac{n_1 k(1-\alpha-\beta)(1-\theta-k\theta)(1-\theta)^{k-1}}{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k} - \frac{n_c k}{1-\theta} + \frac{\sum_{i=1}^{n} c_i x_i}{\theta} \tag{4}$$

Since, $\frac{\partial^2}{\partial\theta^2}\log L$ can be shown to be not in a simple form, therefore a local maximum of L at θ has to be explicitly examined. Finally,

$$\frac{\partial}{\partial k}\log L = -\frac{n_0(1-\alpha-\beta)(1-\theta)^k[\log(1-\theta)]}{\alpha + (1-\alpha-\beta)(1-\theta)^k} + \frac{n_1(1-\alpha-\beta)\theta(1-\theta)^k[1+k\log(1-\theta)]}{\beta + (1-\alpha-\beta)k\theta(1-\theta)^k} + n_c[\log(1-\theta) - \Psi(k)] + \sum_{i=1}^{n} c_i\,\Psi(k+x_i)$$

where Ψ is the digamma function. Since $\frac{\partial^2}{\partial k^2}\log L$ can be shown to be not in a simple form, therefore, a local maximum of L at k has to be explicitly examined. Hence, $\hat{k}$ can be obtained by solving;

$$A(\hat{k}) = 0$$

using any numerical procedure, say, Newton Rapson method, with initial, as given by Alshkaki (2017), where;

$$A(k) = -\frac{n_0(1-\alpha-\beta)(1-\theta)^k[\log(1-\theta)]}{\alpha+(1-\alpha-\beta)(1-\theta)^k} + \frac{n_1(1-\alpha-\beta)\theta(1-\theta)^k[1+k\log(1-\theta)]}{\beta+(1-\alpha-\beta)k\theta(1-\theta)^k} + n_c[\log(1-\theta)-\Psi(k)] + \sum_{i=1}^{n} c_i\,\Psi(k+x_i)$$

Now, letting $\frac{\partial}{\partial\alpha}\log L = 0$, we have from (3) that;

$$1-\alpha-\beta = \frac{n_c}{\frac{n_0}{p_0}[1-(1-\theta)^k] - \frac{n_1}{p_1}[k\theta(1-\theta)^k]}$$

where,

$$p_0 = \alpha + (1-\alpha-\beta)(1-\theta)^k \quad (5)$$

and

$$p_1 = \beta + k(1-\alpha-\beta)\,\theta\,(1-\theta)^k \quad (6)$$

Setting $\frac{\partial}{\partial\theta}\log L = 0$, then (4) reduces to;

$$k(1-\alpha-\beta)(1-\theta)^{k-1}\left[\frac{n_0}{p_0} - \frac{n_1}{p_1}(1-\theta-k\theta)\right] = \frac{\sum_{i=1}^{n} c_i x_i}{\theta} - k\frac{n_c}{1-\theta} \quad (7)$$

Now, if we replace, p_0 and p_1 by their sample relative frequencies, i.e. by their sample estimates, the proportion of zeros and the proportion of ones in the sample, that is; $\widehat{p_0} = n_0/n$ and $\widehat{p_1} = n_1/n$, respectively, then (7) reduce to;

$$nk(k+1)(1-\alpha-\beta)\theta(1-\theta)^{k-1} = \frac{\sum_{i=1}^{n} c_i x_i}{\theta} - k\frac{n_c}{1-\theta} \quad (8)$$

Since the left side of (8), with the use of (2), can be written as;

$$nk(k+1)(1-\alpha-\beta)\theta(1-\theta)^{k-1} = \frac{2n}{\theta(1-\theta)}p_2 \quad (9)$$

Therefore, (8) reduces to;

$$\frac{2n}{\theta(1-\theta)}p_2 = \frac{\sum_{i=1}^{n} c_i x_i}{\theta} - k\frac{n_c}{1-\theta} \quad (10)$$

Now, using the sample relative frequency to estimate p_2, hence, from (10) we have that;

$$\frac{2n_2}{\theta(1-\theta)} = \frac{\sum_{i=1}^{n} c_i\,x_i}{\theta} - k\frac{n_c}{1-\theta}$$

or equivalently, in the form after multiply both sides by $\theta(1-\theta)$;

$$2n_2 = (1-\theta)\sum_{i=1}^{n} c_i\,x_i - \theta k n_c$$

from which we have that;

$$\hat{\theta} = \frac{\sum_{i=1}^{n} c_i\,x_i - 2n_2}{\sum_{i=1}^{n} c_i\,x_i + kn_c}$$

Thus, the estimates of α and β, using the sample relative frequencies estimates, are given by solving (5) and (6) to be;

$$\hat{\alpha} = \frac{(1-\widehat{p_1})(1-\hat{\theta})^{\hat{k}} - \widehat{p_0}\left[1-\hat{k}\hat{\theta}(1-\hat{\theta})^{\hat{k}}\right]}{(1+\hat{k}\hat{\theta})(1-\hat{\theta})^{\hat{k}} - 1}$$

and

$$\hat{\beta} = \frac{\widehat{p_1} - \hat{k}\hat{\theta}(1-\hat{\alpha})(1-\hat{\theta})^{\hat{k}}}{1-\hat{k}\hat{\theta}(1-\hat{\theta})^{\hat{k}}}$$

4. A Simulation Study

In order to check the accuracy of the proposed combined estimation method, we simulated data from different ZOINBD models data sets, then the performance of the estimators are computed through their mean squares errors (MSE) using different sample sizes.

We have used Absoft Pro Fortran compiler for computing, Mathematica and STATISTICA for the needed graphics and other statistical computing. The procedure steps are given below;

(1) Six different ZOINBD models are considered.
(2) Five sample sizes; 15, 30, 50, 100, and 300 are used.
(3) For each sample size, 5,000 random variates were generate from each of the given ZOINBD model.
(4) For each sample size and for each ZOINBD model, the parameters were estimated using the proposed combined estimation method.
(5) The means, standard deviation (SD), bias, and MSE for each of the parameters were computed for each random sample for each sample size of the given ZOINBD models.

Table 1 presents the 6 different simulated ZOINBD Data Sets that were considered, and Tables 2, 3 and 4, represent the findings of the computations.

Table 1. Simulated ZOINBD Data Sets

Data set	Parameters			
	θ	k	α	β
1	0.55	2	0.3	0.1
2	0.4	3	0.5	0.2
3	0.5	4	0.2	0.3
4	0.35	5	0.2	0.1
5	0.2	9	0.5	0.3
6	0.3	15	0.25	0.15

Table 2. Computation Results of Data Sets 1 and 2

n	Parameter	Data Set 1 (0.55, 2. 0.3, 0.1)				Data Set 2 (0.4, 3, 0.5, 0.2)			
		Mean	S.D.	Bias	MSE	Mean	S.D.	Bias	MSE
15	θ	0.672806	0.931367	-0.122806	0.882526	0.466674	0.637911	-0.066674	0.411376
	k	0.93098	1.525467	1.069020	3.469853	2.10333	3.754639	0.896670	14.901331
	α	0.134868	0.354564	0.165132	0.152984	0.475856	0.512556	0.024144	0.263297
	β	0.054113	0.434129	0.045887	0.190574	0.190455	0.621534	0.009545	0.386396
30	θ	0.58132	0.810213	-0.031320	0.657426	0.442126	0.544256	-0.042126	0.297989
	k	1.68711	1.332417	0.312890	1.873235	2.4048	3.198456	0.595200	10.584384
	α	0.275601	0.343345	0.024399	0.118481	0.486322	0.493133	0.013678	0.243367
	β	0.090557	0.413423	0.009443	0.171008	0.194321	0.6107998	0.005679	0.373109
50	θ	0.560331	0.515333	-0.010331	0.265675	0.408846	0.498322	-0.008846	0.248403
	k	1.89541	1.132434	0.104590	1.293346	2.87132	2.365454	0.128680	5.611931
	α	0.293063	0.324911	0.006937	0.105615	0.497806	0.421134	0.002194	0.177359
	β	0.097243	0.382125	0.002757	0.146027	0.199114	0.5777899	0.000886	0.333842
100	θ	0.55748	0.431522	-0.007480	0.186267	0.404926	0.367911	-0.004926	0.135383
	k	1.92491	1.036767	0.075090	1.080524	2.93132	2.167667	0.068680	4.703497
	α	0.295194	0.303578	0.004806	0.092183	0.498991	0.3745667	0.001009	0.140301
	β	0.098116	0.351796	0.001884	0.123764	0.199647	0.521534	0.000353	0.271998
300	θ	0.551554	0.375793	-0.001554	0.141223	0.400608	0.355799	-0.000608	0.126593
	k	1.98721	0.983241	0.012790	0.966926	2.99877	1.537867	0.001230	2.365036
	α	0.299462	0.297122	0.000538	0.088282	0.500261	0.3433456	-0.000261	0.117886
	β	0.099889	0.343945	0.000111	0.118298	0.200228	0.441722	-0.000228	0.195118

Table 3. Computation Results of Data Sets 3 and 4

n	Parameter	Data Set 3 (0.5, 4, 0.2, 0.3)				Data Set 4 (0.35, 5, 0.2, 0.1)			
		Mean	S.D.	Bias	MSE	Mean	S.D.	Bias	MSE
15	θ	0.413318	0.983256	0.086682	0.974306	0.313642	0.632145	0.036358	0.400929
	k	5.83731	3.643545	-1.837310	16.651128	6.01145	4.066667	-1.011450	17.560812
	α	0.20995	0.632133	-0.009950	0.399691	0.210303	0.672778	-0.010303	0.452736
	β	0.311111	0.625589	-0.011111	0.391485	0.108426	0.524344	-0.008426	0.275008
30	θ	0.450354	0.771323	0.049646	0.597404	0.322709	0.533279	0.027291	0.285131
	k	4.96511	3.356456	-0.965110	12.197234	5.73971	3.655323	-0.739710	13.908557
	α	0.206305	0.503145	-0.006305	0.253195	0.207983	0.544223	-0.007983	0.296242
	β	0.306721	0.541733	-0.006721	0.293520	0.106483	0.422412	-0.006483	0.178474
50	θ	0.474479	0.509127	0.025521	0.259862	0.34145	0.499799	0.008550	0.249872
	k	4.47021	3.156578	-0.470210	10.185082	5.23171	3.373434	-0.231710	11.433746
	α	0.203466	0.473717	-0.003466	0.224420	0.202922	0.481442	-0.002922	0.231795
	β	0.303557	0.489127	-0.003557	0.239258	0.102392	0.399774	-0.002392	0.159825
100	θ	0.494148	0.424317	0.005852	0.180079	0.345708	0.407133	0.004292	0.165776
	k	4.10249	2.965767	-0.102490	8.806278	5.11434	3.137656	-0.114340	9.857959
	α	0.200819	0.393411	-0.000819	0.154773	0.01592	0.403321	0.184080	0.196553
	β	0.30081	0.401434	-0.000810	0.161150	0.101349	0.332945	-0.001349	0.110854
300	θ	0.497811	0.357678	0.002189	0.127938	0.345831	0.366213	0.004169	0.134129
	k	4.03721	2.454578	-0.037210	6.026338	5.11122	2.554546	-0.111220	6.538075
	α	0.200289	0.377189	-0.000289	0.142272	0.201556	0.387991	-0.001556	0.150539
	β	0.300275	0.354565	-0.000275	0.125716	0.101321	0.312667	-0.001321	0.097762

Table 4. Computation Results of Data Sets 5 and 6

n	Parameter	Data Set 5 (0.2, 9, 0.4, 0.3)				Data Set 6 (0.3, 15, 0.25, 0.15)			
		Mean	S.D.	Bias	MSE	Mean	S.D.	Bias	MSE
15	θ	0.142279	0.705341	0.057721	0.500838	0.268572	0.757723	0.031428	0.575132
	k	13.9758	6.065795	-4.975800	61.552455	17.52147	7.145345	-2.521470	57.413766
	α	0.406544	0.646678	-0.006544	0.418235	0.250353	0.686543	-0.000353	0.471341
	β	0.305425	0.652341	-0.005425	0.425578	0.151073	0.593225	-0.001073	0.351917
30	θ	0.166064	0.635467	0.033936	0.404970	0.27577	0.717736	0.024230	0.515732
	k	11.51251	5.198789	-2.512510	33.340114	16.89241	6.653379	-1.892410	47.848668
	α	0.404104	0.579978	-0.004104	0.336391	0.250277	0.555211	-0.000277	0.308259
	β	0.303345	0.417223	-0.003345	0.174086	0.150835	0.542745	-0.000835	0.294573
50	θ	0.172202	0.507778	0.027798	0.258611	0.284958	0.621178	0.015042	0.386088
	k	10.98733	4.665343	-1.987330	25.714906	16.13565	6.376545	-1.135650	41.950027
	α	0.403433	0.499789	-0.003433	0.249801	0.250175	0.498999	-0.000175	0.249000
	β	0.30279	0.398889	-0.002790	0.159120	0.150522	0.447667	-0.000522	0.200406
100	θ	0.188987	0.445322	0.011013	0.198433	0.29408	0.515444	0.005920	0.265718
	k	9.72531	4.094533	-0.725310	17.291275	15.43113	5.338779	-0.431130	28.688434
	α	0.401503	0.431567	-0.001503	0.186252	0.250069	0.534543	-0.000069	0.196759
	β	0.301235	0.375333	-0.001235	0.140876	0.150201	0.400337	-0.000201	0.160270
300	θ	0.195361	0.388178	0.004639	0.150704	0.298267	0.443572	0.001733	0.196759
	k	9.30291	3.938767	-0.302910	15.605640	15.12222	4.976889	-0.122220	24.784362
	α	0.400732	0.397778	-0.000732	0.158228	0.250018	0.397845	-0.000018	0.158281
	β	0.300629	0.372355	-0.000629	0.138649	0.150051	0.3875667	-0.000051	0.150208

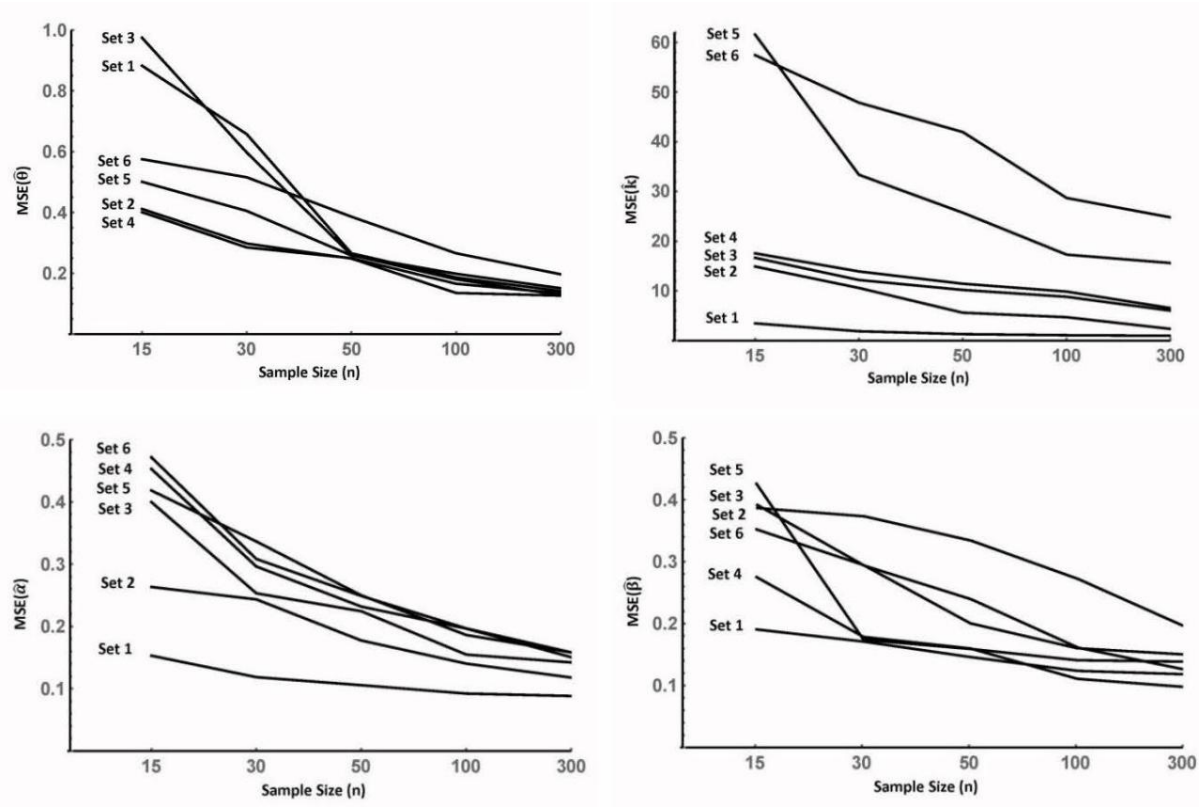

Figure 1. Influence of the sample sizes on the estimated parameter's MSE for the 6 simulated data Sets

Figure 1 presents the plots of the MSE of each of the parameter estimates for the 6 simulated data set, which shows that the MSE of each of the estimated parameter for each of the different ZOINBD models is decreasing as the sample size increases.

5. Fitting Zero-One Inflated Negative Binomial Distributions to Real Life Data

We have taken six real life data set from different filed in order to show the usefulness of the proposed combined estimation procedure to estimate and fit ZOINBD to these real life data sets. The data sets are;

Data Set 1: Represents the number of units of consumers good purchased by households over 26 weeks, see Lindsey (1995). This data was studied by Aryuyuen et al. (2014) using the zero inflated negative binomial-generalized exponential distribution.

Data Set 2: Represents the number of major derogatory reports in the credit history of individual credit card applicants, Greene (1994). This data was studied by Saengthong et al. (2015) using the zero inflated negative binomial – crack distribution.

Data Set 3: Represents the number of Stillbirths in 402 litters of New Zealand white rabbits, Morgan et al. (2007). This data was studied by Morgan et al. (2007) using the zero-inflated Poisson distribution.

Data Set 4: Represents the number of hospital stays by United States residents aged 66 and over, Flynn (2009). This data was studied by Aryuyuen et al. (2014) using the zero inflated negative binomial-generalized exponential distribution.

Data Set 5: Represents the number of households according to the total number of migrants in household cohort excluding international migrants of the rural areas of Comilla district of Bangladesh, Pandey and Tiwari (2011). This data was studied by Pandey and Tiwari (2011) using a mixture of a geometric and log-series distributions.

Data Set 6: Represents the number of migrants from a household in growth-center villages, Pandey and Tiwari (2011). This data was studied by Pandey and Tiwari (2011) using a mixture of a geometric and log-series distributions.

We have chosen these data sets due to the fact that their observed relative frequencies at zeros and ones are noticeably large as can be seen from Table 5, for example, the relative frequency of zeros for Data Set 1 is 0.806, and the for the ones is 0.082, hence both sum to 0.888, which is a noticeable large, and similarly, the same note for the other Data Sets 2 to 6, inducting that the ZOINBD models may be an appropriate model to be considered for these data sets.

Table 5. Observed Relative Frequencies for the Data Sets

X	Data Set					
	1	2	3	4	5	6
0	0.806	0.8036	0.787	0.8037	0.8173	0.8087
1	0.082	0.1039	0.1203	0.136	0.1229	0.1032
Total	0.888	0.9075	0.9073	0.9397	0.9402	0.9119

Table 6. The Proposed Estimation Results for Data Sets 1, 2 and 3

X	Data Set 1 Frequencies		Data Set 2 Frequencies		Data Set 3 Frequencies	
	Observed	Estimated	Observed	Estimated	Observed	Estimated
0	1612	1612	1060	1060	314	314
1	164	164	137	137	48	48
2	71	71	50	50	20	20
3	47	46	24	27	7	9
4	28	31	17	16	5	4
5	17	21	11	9	2	2
6	12	15	5	6	2	1
7	12	11	6	4	1	1
8	5	8	9	10		
9	7	6				
10+	25	15				
Total	2000	2000	1319	1319	399	399
Model Parameters	θ	0.76518		0.7052		0.53744
	k	0.52649		0.3418		0.5472
	α	0.65496		0.4699		0.3851
	β	0.02115		0.0234		0.00211
	χ^2	3.0566		2.1069		1.6944
	df	6		4		3
	p-value	0.8017		0.7161		0.6382

Table 7. The Proposed Estimation Results for Data Sets 4, 5 and 6

X	Data Set 4 Frequencies		Data Set 5 Frequencies		Data Set 6 Frequencies	
	Observed	Estimated	Observed	Estimated	Observed	Estimated
0	3541	3541	1941	1941	972	972
1	599	599	292	292	124	124
2	176	171	67	67	32	32
3	48	59	37	35	25	23
4	20	22	17	19	12	16
5	12	8	6	10	10	11
6	5	3	7	5	5	7
7	1	1	3	3	5	5
8+	4	2	5	3	17	12
Total	4406	4406	2375	2375	1202	1202
Model Parameters	θ	0.43487		0.55372		0.62853
	k	0.4013		0.8272		1.4815
	α	0.07502		0.69552		0.77199
	β	0.00879		0.06718		0.06903
	χ^2	7.7122		4.0582		3.9196
	df	4		4		4
	p-value	0.1027		0.3982		0.417

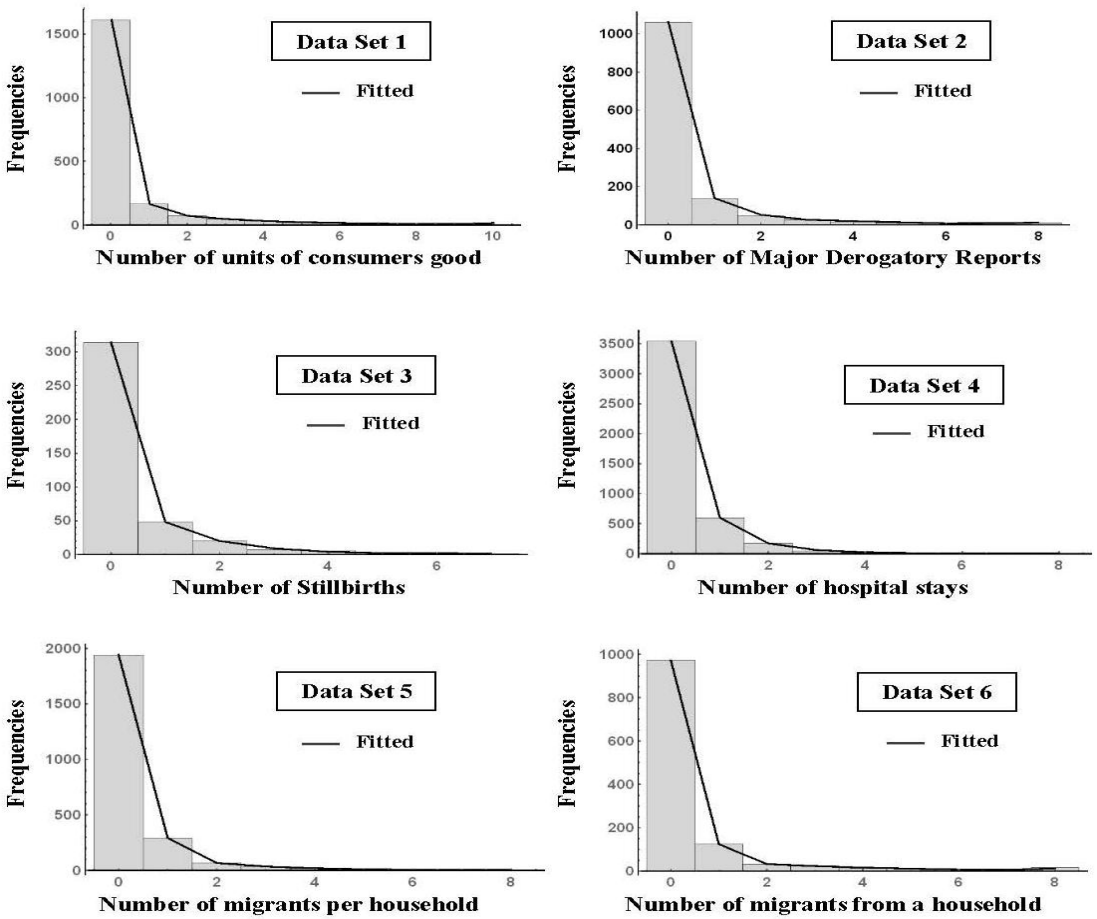

Figure 2. Observed and Estimated Frequencies of All Data Sets

Our proposed estimation procedures was used to estimate the parameters of the ZOINBD model using each of the six data sets. Tables 6 and 7 show the observed and estimated frequencies, the estimated parameters, and the chi-squares goodness of fit test for each of the data sets. From these results, our proposed estimation procedures gives good estimates statistically. These results can be seen visually also from Figure 2, illustrating the graphs of the distributions of the observed and estimated frequencies for each of the data sets.

6. Conclusions

We considered estimation of the parameters of the zero-one inflated negative binomial distribution by a combined method of relative frequencies and maximum likelihood estimators. We simulated six different zero-one inflated negative binomial distribution models data sets, in order to check the performance of the proposed estimation method, and the mean squares errors of each of the estimated parameter was computed using different sample sizes. The mean squares error of each of the estimated parameter for each of the six simulated data shows that the it is decreasing as the sample size increases. We used the proposed estimation procedures to estimate the parameters of the zero-one inflated negative binomial distribution model of six different real life data sets, and it gave a good results visually, supported by the results of the chi-squares goodness of fit test for each of the data sets.

REFERENCES

[1] Alshkaki, R. S. (2017). Moments Estimators of the Parameters of the Zero-One Inflated Negative Binomial Distribution, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, Vol:11, No:1, 38-41.

[2] Aryuyuen, S., Bodhisuwan, W. and Supapakorn, T. (2014). Zero inflated negative binomial-generalized exponential distribution and its applications, Songklanakarin J. Sci. Technol, 36 (4), 483-491.

[3] Astuti, C. C., and Mulyanto, A. D., Estimation Parameters And Modelling Zero Inflated Negative Binomial. CAUCHY – JURNAL MATEMATIKA MURNI DAN APLIKASI Volume 4(3) (2016), Pages 115-119. DOI: 10.18860/ca.v4i3.3656.

[4] Flynn, M. (2009). More flexible GLMs zero-inflated models and hybrid models. Casualty Actuarial Society EForum, Winter, U.S.A., 148-224.

[5] Gan, N., General Zero-Inflated Models and Their Applications. PhD Thesis, NC State University, 2000.

[6] Greene, W. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, New York University, New York, U.S.A.

[7] Johnson, N. L., Kemp, A. W., and and Kotz, S. (2005), Univariate Discrete Distributions, Third Edition, John Wiley and Sons, New Jersey.

[8] Lindsey, J. K. 1995. Modelling Frequency and Count Data. Oxford science publications, Clarendon Press, UK., p. 146.

[9] Lukusa, T. M., Lee, S. M., and Li, C. S., Review of Zero-Inflated Models with Missing Data. Current Research in Biostatistics, Volume 7, Issue 1, 2017, 1-12. DOI:10.3844/amjbsp.2017.1.12.

[10] Morgan, B. J. T., Palmer, K. J., and Ridout, M. S. (2007). Score Test Oddities. The American Statistician, 61: 285–288.

[11] Pandey, H. and Tiwari, R. (2011), An Inflated Probability Model for the Rural Out-Migration, Recent Research in Science and Technology 2011, 3(7): 100-103.

[12] Phang, Y. N. and Loh, E. F. Zero Inflated Models for Overdispersed Count Data. World Academy of Science, Engineering and Technology, International Journal of Health and Medical Engineering, Vol:7, No:8, 2013, 1331-1333.

[13] Preisser, J. S., Stamm, J. W., Long, D. L. and Kincade, M. E., Review and Recommendations for Zero-Inflated Count Regression Modeling of Dental Caries Indices in Epidemiological Studies. Caries Res 2012; 46: 413–423. https://doi.org/10.1159/000338992.

[14] Saengthong, P., Bodhisuwan, W., and Thongteeraparp, A. (2015). The Zero Inflated Negative Binomial – Crack Distribution: Some Properties And Parameter Estimation. Songklanakarin J. Sci. Technol, 37(6), 701-711.

[15] Staub, K. E. and Winkelmann, R., Consistent Estimation of Zero-Inflated Count Models. HEALTH ECONOMICS (2012), Wiley Online Library. DOI: 10.1002/hec.2844.

[16] Yang, S., Harlow, L. L., Puggioni, G., and Redding, C. A., A Comparison of Different Methods of Zero-Inflated Data Analysis and an Application in Health Surveys, Journal of Modern Applied Statistical Methods, May 2017, Vol. 16, No. 1, 518-543. doi: 10.22237/jmasm/1493598600.

[17] Yusuf O, Bello T, and Gureje O. Zero Inflated Poisson and Zero Inflated Negative Binomial Models with Application to Number of Falls in the Elderly. Biostat Biometrics Open Acc J. 2017; 1(4): 555-566. DOI:10.19080/BBOAJ.2017. 01.555566.

[18] Zamri, N. S. N. and Zamzuri, Z. H., A review on models for count data with extra zeros, AIP Conference Proceedings 1830, 080010 (2017); https://doi.org/10.1063/1.4980994.