

## Estimation in zero-inflated Generalized Poisson distribution

**Kirtee K. Kamalja and Yogita S. Wagh**

*Department of Statistics, School of Mathematical Sciences,  
North Maharashtra University, Jalgaon, India  
kirteekamalja@gmail.com, yvj.jalasri@gmail.com*

### Abstract

Overdispersion is a common phenomenon in Poisson modelling. The generalized Poisson (GP) distribution accommodates both overdispersion and under dispersion in count data. In this paper, we briefly overview different overdispersed and zero-inflated regression models. To study the impact of fitting inaccurate model to data simulated from some other model, we simulate data from ZIGP distribution and fit Poisson, Generalized Poisson (GP), Zero-inflated Poisson (ZIP), Zero-inflated Generalized Poisson (ZIGP) and Zero-inflated Negative Binomial (ZINB) model. We compare the performance of the estimates of Poisson, GP, ZIP, ZIGP and ZINB through mean square error, bias and standard error when the samples are generated from ZIGP distribution. We propose estimators of parameters of ZIGP distribution based on the first two sample moments and proportion of zeros referred to as MOZE estimator and compare its performance with maximum likelihood estimate (MLE) through a simulation study. It is observed that MOZE are almost equal or even more efficient than that of MLE of the parameters of ZIGP distribution.

**Keywords:** Zero-inflated Generalized Poisson, Zero-inflated Negative Binomial, simulation.

**Mathematics subject classification:** 65Cxx, 62F10, 62J02, 97K80

---

### 1. Introduction

Statistical models that address the count data have been implemented in many areas such as insurance, dental epidemiology, health care facilities, risk classification, medicine, etc. The Poisson model is a standard approach to analyze the count data. The most annoying property of Poisson model is equality of mean and variance, which is known as equidispersion. In real practice, the sample data often has variance which exceeds its mean. The phenomenon of excess variability is called overdispersion and has been widely studied in the literature (Dean and Lawless (1989) and Dean (1992)). Failure to properly address existing overdispersion leads to serious underestimation of standard errors and misleading inference for the regression

parameters. Consequently, several models and associated estimation methods have been proposed for handling overdispersed data. These models include the two popular versions of negative binomial (NB) model referred to as the NB-1 and NB-2 regression model and are discussed by McCullagh and Nelder (1989).

Other common distribution to handle overdispersion/under dispersion is generalized Poisson distribution (GPD) introduced by Consul and Jain (1973). The classical GP regression model is introduced by Consul (1989) and is referred as the GP-1 model. Further, this model is deeply studied by Consul and Famoye (1992) and used by Wulu, Singh, Famoye, McGwin (2002) to analyse farm injury data. Famoye (1993) introduced restricted GPR model and is referred to as a GP-2 regression model. The GP-1 and GP-2 regression models are the natural extensions of the Poisson regression model. Thus, the NB and GP model are used to handle dispersion in the data.

Another cause of overdispersion is excess zeroes in the data which is detected when the frequency of 'zero' observation is significantly higher than the one predicted by the Poisson model. Neyman (1939) followed by Feller (1945) introduces the concept of zero-inflation where the data contains more zeros. This concept leads to zero-inflated versions of basic distributions such as ZIP, ZINB. Particularly, the zero-inflated Poisson (ZIP) model is not appropriate to handle the issue of excess zeros with overdispersion. To deal with such a data, the most commonly used zero-inflated model is zero-inflated negative binomial (ZINB) model, which is a mixture of distribution degenerate at zero with a baseline negative binomial distribution. This model is used to fit overdispersed data with zero-inflation and has been discussed in detail by Cameron and Trivedi (1986) and Gurmu, Rilstone, Stern (1999).

An alternative to ZINB model is zero-inflated generalized Poisson (ZIGP) model, which is a mixture of distribution degenerate at zero with a baseline GPD. The ZIGP regression model is applied in different fields to model zero-inflated and overdispersed count data. Some of the studies include analysis of patent outsourcing data (Czado, Erhardt, Min 2007), the fetal movement data (Gupta, Gupta, Tripathi 1996), an injury and accident related data, motor insurance data (Yip and Yau 2005), domestic violence data (Famoye and Singh 2006), Mosquito count data (Lawal 2012), an auto insurance claim data (Wagh and Kamalja 2017a) etc. Other models in the literature include the hurdle model (Mullahy 1986), the two-part model (Heibron 1994) the semi-parametric model (Gurmu 1997) etc.

Comparison between the GPD and Negative Binomial Distribution (NBD) reveals slight differences in many aspects. The GPD has a heavier tail while the NBD has a larger mass at zero. However, their respective zero-inflated distributions with masses at zero and fixed

mean and variance can differ even more from each other (see Joe and Zhu 2005). There is a relatively unknown number of situations in which the iterative estimation technique for parameters of ZINB regression model fail to converge. While estimation in case of ZIGP model converges more often.

In models used for count data, the assessment of over or under dispersion and zero-inflation is very important. Gupta, Gupta and Tripathi (2004) developed a score test to assess both zero-inflation and dispersion in a ZIGP model. Xiang, Lee, Yau and McLachlan (2007) proposed a score test for assessing overdispersion in the ZIP mixed regression model against a ZINB mixed alternative. Yang, Hardin and Addy (2009) proved that the score statistic for testing overdispersion in ZIP against the ZIGP is the same as that of the score statistic for testing overdispersion in ZIP against ZINB. Yang, Hardin and Addy (2010) developed the score statistic for testing zero-inflation with respect to two forms of GP model. Zamani and Ismail (2013) proposed the score test for testing overdispersion in the ZIP model against ZIGP alternatives.

When overdispersion within zero-inflated data are ignored, it results in poor estimation and inflated Type I error. Failure to accurately model existing overdispersion leads to severe underestimation of standard errors and misleading results for regression parameters (Yang, Hardin and Addy 2009). Chaney, Charity, David, and Aban (2013) simulated data from ZIP/ZINB distribution and fitted Poisson, NB, ZIP, and ZINB models and studied the consequences of misspecifying the statistical model.

In most of the research problems, it has been observed that the ZIP regression model is inadequate to model excess zeros and the ZINB regression model could not be a suitable fit to such a count data. The major problem is non-convergence of iterative technique to estimate the parameters of ZINB model. Such models do not successfully fit the data. The alternative to ZIP and ZINB model is ZIGP model. We motivated to study the effect of fitting different count data models to the data simulated from ZIGP model. Further, we aim to propose new estimators of parameters of ZIGP model which would be simple in terms of calculation and estimation can be done without specific software. This lead to the new estimator MOZE. The paper is organized as follows.

In section 2, we briefly overview different GP models and their respective zero-inflated models, which are most common for modeling overdispersed and zero-inflated count data. We present a simulation study in Section 3 for choosing between standard zero-inflated models by comparing the performance of the estimates of Poisson, GPD, ZIP, ZIGP and ZINB models when the data come from ZIGP distribution. In Section 4, we demonstrate the behavior of the

zero-inflation index with respect to the inflation parameter. In Section 5, we propose estimator of parameters of ZIGP distribution based on moments and the proportion of zeros and compare their performances with respective MLEs through a simulation study in Section 6. In Section 7, we present a numerical study in which different count data models are fitted to some real-life datasets.

## 2. An overview of overdispersed and zero-inflated count data models

Several forms of GP model have been introduced by the researchers. The primary interest to introduce the different functional forms of GP model is to use the proper mean-variance interrelation in a regression context. The other objective is to achieve flexibility in the development of various test procedures and inferential advantages.

Consul and Jain (1973), Consul (1989), Consul and Famoye (1992) and Famoye (1993) introduced various forms of GP model and studied these models. Zamani and Ismail (2012) introduced a new functional form of the GP regression model, which is referred to as the GP-P model, that parametrically nests the Poisson and two well-known GP regression models (GP-1 and GP-2). We summarize the functional forms of GP regression model in Table 2.1

If the count data contains excess zeros along with overdispersion, one may consider ZIGP or ZINB model to fit the data. Parallel to the functional forms of GP models, functional forms of ZIGP models are also generated. Further, Zamani and Ismail (2014) proposed a ZIGP regression model, which mixes a distribution degenerate at zero along with the GP-P distribution. The ZIGP-1 and ZIGP-2 regression models are particular cases of ZIGP-P model with  $P = 1$  and  $P = 2$  respectively. The functional forms of ZIGP regression models are summarized in Table 2.2.

Table 2.1 Different Poisson based regression models and their mean and variance

Sr. No.	Poisson and different types of GPR model with introducer	Parameter of the regression model	pmf of Y: $P(Y_i = y_i   \text{parameters}), y_i = 0, 1, 2, \dots, \theta_i > 0$	(Mean, variance)
1.	Poisson (Frome, Kutner, Beauchamp (1973))	$\theta_i$	$\frac{\theta_i^{y_i}}{y_i!} \exp(-\theta_i)$	$E(Y_i) = \theta_i$ $V(Y_i) = \theta_i$
2.	GP-1 (Consul and Jain (1973))	$(\theta_i, \varphi)$	$\frac{\theta_i(\theta_i + \varphi y_i)^{y_i-1}}{(1 + \varphi)^{y_i} y_i!} \exp\left(-\frac{\theta_i + \varphi y_i}{1 + \varphi}\right)$	$E(Y_i) = \theta_i$ $V(Y_i) = \theta_i(1 + \varphi)^2$ Here $\varphi$ is dispersion parameter
3.	GP-1 (Consul (1989))	$(\theta_i, \varphi)$	$\frac{((1 - \varphi)\theta_i + \varphi y_i)^{y_i-1} (1 - \varphi)\theta_i}{y_i!} \exp(-((1 - \varphi)\theta_i - \varphi y_i))$	$E(Y_i) = \theta_i$ $V(Y_i) = \varphi E(Y_i)$ Here $\varphi = \frac{1}{(1 - \varphi)^2}$ is dispersion factor
4.	GP-1 (Consul (1989))	$(\theta_i, \varphi)$	$\frac{\theta_i(\theta_i + \varphi y_i)^{y_i-1}}{y_i!} \exp(-\theta_i - \varphi y_i), \max\left(-1, -\frac{\theta_i}{4}\right) < \varphi < 1$	$E(Y_i) = \frac{\theta_i}{(1 - \varphi)}$ $V(Y_i) = \varphi E(Y_i)$ Here $\varphi = \frac{1}{(1 - \varphi)^2}$ is dispersion factor
5.	GP (Consul and Famoye (1992))	$(\theta_i, \varphi)$	$\frac{\theta_i(\theta_i + y_i(\varphi - 1))^{y_i-1} \varphi^{-y_i}}{y_i!} \exp(-(\theta_i + y_i(\varphi - 1)) / \varphi)$	$E(Y_i) = \theta_i$ $V(Y_i) = \theta_i \varphi^2$
6.	GP-2 (Famoye (1993))	$(\theta_i, \varphi)$	$\left(\frac{\theta_i}{1 + \varphi \theta_i}\right)^{y_i} \frac{(1 + \varphi y_i)^{y_i-1}}{y_i!} \exp\left(-\frac{\theta_i(1 + \varphi y_i)}{1 + \varphi \theta_i}\right)$	$E(Y_i) = \theta_i$ $V(Y_i) = \theta_i(1 + \varphi \theta_i)^2$
7.	GP-P (Zamani and Ismail (2012))	$(\theta_i, \varphi, P)$	$\frac{\theta_i(\theta_i + \varphi \theta_i^{P-1} y_i)^{y_i-1}}{(1 + \varphi \theta_i^{P-1} y_i)^{y_i}} \exp\left(-\frac{\theta_i + \varphi \theta_i^{P-1} y_i}{1 + \varphi \theta_i^{P-1} y_i}\right)$	$E(Y_i) = \theta_i$ $V(Y_i) = \theta_i(1 + \varphi \theta_i^{P-1})^2$

Table 2.2 Various forms of zero-inflated Poisson based regression models with their mean and variance

Sr. No.	Type of Zero-Inflated regression model and introducer/user	Parameters of the regression model	pmf of Y: $P(Y_i = y_i   \text{parameters}), y_i = 0, 1, 2, \dots$ $\theta_i > 0, 0 \leq \omega_i < 1$	(Mean, variance)	Remark
1.	ZIP (Lambert (1992))	$(\theta_i, \omega_i)$	$\begin{cases} \omega_i + (1 - \omega_i)e^{-\theta_i}, & y_i = 0 \\ (1 - \omega_i) \frac{\theta_i^{y_i}}{y_i!} e^{-\theta_i}, & y_i > 0 \end{cases}$	$\begin{cases} E(Y_i) = (1 - \omega_i)\theta_i \\ V(Y_i) = (1 - \omega_i)\theta_i(1 + \omega_i\theta_i) \end{cases}$	Reduces to Poisson for $\omega_i = 0$
2.	ZIGP (Czado and Min (2005))	$(\theta_i, \phi, \omega_i)$	$\begin{cases} \omega_i + (1 - \omega_i)e^{-\theta_i/\phi}, & y_i = 0 \\ (1 - \omega_i)\theta_i(\theta_i + y_i(\phi - 1))^{y_i-1} \frac{\phi^{-y_i}}{y_i!} \exp\left(-\frac{\theta_i + y_i(\phi - 1)}{\phi}\right), & y_i > 0 \end{cases}$	$\begin{cases} E(Y_i) = (1 - \omega_i)\theta_i \\ V(Y_i) = (1 - \omega_i)\theta_i(\phi^2 + \omega_i\theta_i) \end{cases}$	Reduces to ZIP for $\phi = 1$ (Here $\phi$ is dispersion parameter)
3.	ZIGP-1 (Yang, Harčin and Abby (2009))	$(\theta_i, \phi, \omega_i)$	$\begin{cases} \omega_i + (1 - \omega_i)e^{-(1-\phi)\theta_i}, & y_i = 0 \\ (1 - \omega_i)(1 - \phi)\theta_i + \phi y_i)^{y_i-1} \frac{(1 - \phi)\theta_i}{y_i!} \exp(-(1 - \phi)\theta_i - \phi y_i), & y_i > 0 \end{cases}$	$\begin{cases} E(Y_i) = (1 - \omega_i)\theta_i \\ V(Y_i) = (1 - \omega_i)\theta_i \left( \frac{1}{(1 - \phi)^2} + \omega_i\theta_i \right) \end{cases}$	Reduces to ZIP for $\phi = 0$ (Here $\frac{1}{(1 - \phi)^2}$ is dispersion factor)
4.	ZIGP-2 (Yang, Harčin and Abby (2009))	$(\theta_i, \phi, \omega_i)$	$\begin{cases} \omega_i + (1 - \omega_i) \exp\left(-\left(\frac{\theta_i}{1 + \phi}\right)\right), & y_i = 0 \\ (1 - \omega_i) \left(\frac{\theta_i}{1 + \phi}\right)^{y_i} \frac{(1 + \phi)^{y_i-1}}{y_i!} \exp\left(-\frac{\theta_i(1 + \phi)}{1 + \phi}\right), & y_i > 0 \end{cases}$	$\begin{cases} E(Y_i) = (1 - \omega_i)\theta_i \\ V(Y_i) = (1 - \omega_i)\theta_i(1 + \phi\theta_i)^2 + \omega_i\theta_i \end{cases}$	Reduces to ZIP for $\phi = 0$ based on GP-2 (Famoye (1993))
5.	ZIGP-1 (Zamani and Ismail (2014))	$(\theta_i, \phi, \omega_i)$	$\begin{cases} \omega_i + (1 - \omega_i) \exp\left(\frac{-\theta_i}{1 + \phi}\right), & y_i = 0 \\ (1 - \omega_i) \frac{\theta_i(\theta_i + \phi)^{y_i-1}}{(1 + \phi)^{y_i}} \exp\left(-\frac{\theta_i + \phi y_i}{1 + \phi}\right), & y_i > 0 \end{cases}$	$\begin{cases} E(Y_i) = (1 - \omega_i)\theta_i \\ V(Y_i) = (1 - \omega_i)\theta_i(1 + \phi)^2 + \omega_i\theta_i \end{cases}$	Reduces to ZIP for $\phi = 0$
6.	ZIGP-P (Zamani and Ismail (2014))	$(\theta_i, \phi, \omega_i, P)$	$\begin{cases} \omega_i + (1 - \omega_i) \exp\left(\frac{-\theta_i}{1 + \phi P^{P-1}}\right), & y_i = 0 \\ (1 - \omega_i) \frac{\theta_i(\theta_i + \phi P^{P-1})^{y_i-1}}{(1 + \phi P^{P-1})^{y_i}} \exp\left(\frac{-\theta_i + \phi P^{P-1} y_i}{1 + \phi P^{P-1}}\right), & y_i > 0 \end{cases}$	$\begin{cases} E(Y_i) = (1 - \omega_i)\theta_i \\ V(Y_i) = (1 - \omega_i)\theta_i \times \left( (1 + \phi P^{P-1})^2 + \omega_i\theta_i \right) \end{cases}$	Reduces to ZIGP-1 for $P = 1$ and ZIGP-2 for $P = 2$

### 3. A Simulation Study

In this section, we perform a simulation study to perceive the impact of choosing an inappropriate model on MLE, given that the data is generated from ZIGP distribution. We present the results of a simulation study, which fits Poisson (P), GP, ZIP, ZIGP and ZINB model to the sample data from ZIGP distribution. We compare the MSE, bias and standard error (SE) of the mean parameter for these fits. We observe a clear picture of the poor performance of the estimates of the mean parameter when the overdispersion and/or zero-inflation in a simulated data from ZIGP is ignored. We briefly discuss the methodology to estimate the mean parameter and give a systematic algorithm for a simulation study performed.

#### 3.1 Algorithm for comparison of performance of mean parameter estimate

Let  $y_1, y_2, \dots, y_n$  be a random sample from ZIGP distribution. To fit Poisson, GP, ZIP, ZIGP and ZINB distribution to this data, we use ZIGP, VGAM and *pscl* packages in R, which actually fit, the intercept model and estimate the intercept parameter  $\beta_0$  of the corresponding regression model. Let  $\hat{\theta}^d$  be the MLE of  $\theta$  (mean parameter) when model  $d$  is fitted to the random sample,  $d = P, GP, ZIP, ZIGP, ZINB$ . Then  $\hat{\theta}^d$  is estimated as  $\hat{\theta}^d = e^{\hat{\beta}_0}$ . The systematic algorithm for comparison of the performance of estimate of mean parameter of the different distributions is given below.

Generate  $m = 500$  random samples each of size  $n$  from  $ZIGP(\theta, \varphi, \omega) |_{\theta=2, \varphi=1.5, \omega=\omega_0}$ , distribution for zero-inflation parameter  $\omega_0, \omega_0 = 0.1, 0.2, \dots, 0.8$ .

- i) Fit Poisson, GP, ZIP, ZIGP and ZINB models to the  $i^{th}$  random sample and obtain MLE of  $\theta$  as  $\hat{\theta}_i^P, \hat{\theta}_i^{GP}, \hat{\theta}_i^{ZIP}, \hat{\theta}_i^{ZIGP}, \hat{\theta}_i^{ZINB}$  respectively for  $i = 1, 2, \dots, m$ .
- ii) Evaluate Mean square error (MSE) and bias of  $\hat{\theta}^d$  as  $MSE(\hat{\theta}^d), bias(\hat{\theta}^d)$  using following formulae.

$$MSE(\hat{\theta}^d) = \frac{\sum_{i=1}^m (\hat{\theta}_i^d - \theta)^2}{m} \text{ and } bias(\hat{\theta}^d) = \frac{\sum_{i=1}^m (\hat{\theta}_i^d - \theta)}{m}, d = P, GP, ZIP, ZIGP, ZINB.$$

- iii) Obtain  $SE(\hat{\theta}^d) = \sqrt{var(\hat{\theta}^d)}$  where  $var(\hat{\theta}^d) = \frac{\sum_{i=1}^m (\hat{\theta}_i^d - \bar{\theta}^d)^2}{m-1}$  and  $\bar{\theta}^d = \frac{\sum_{i=1}^m \hat{\theta}_i^d}{m}$ .

- iv) Repeat Steps (i), (ii) and (iii) across different values of  $\omega_0 = 0.1, 0.2, \dots, 0.8$  to get the  $MSE(\hat{\theta}^d), bias(\hat{\theta}^d)$  and  $SE(\hat{\theta}^d)$  for each  $\omega_0$ .

- v) Plot  $MSE(\hat{\theta}^d)|_{\omega=\omega_0}, bias(\hat{\theta}^d)|_{\omega=\omega_0}$  and  $SE(\hat{\theta}^d)|_{\omega=\omega_0}$  versus  $\omega_0 = 0.1, 0.2, \dots, 0.8$ .

The results of a simulation study of random samples of size  $n = 50, 100$  are given in Table 3.1 and Figure 3.1-3.10. To observe the validity of results, a parallel simulation study is performed for  $\theta = 5$  and  $n = 250$  which lead to the similar conclusions. Hence, these results are not included in the paper.

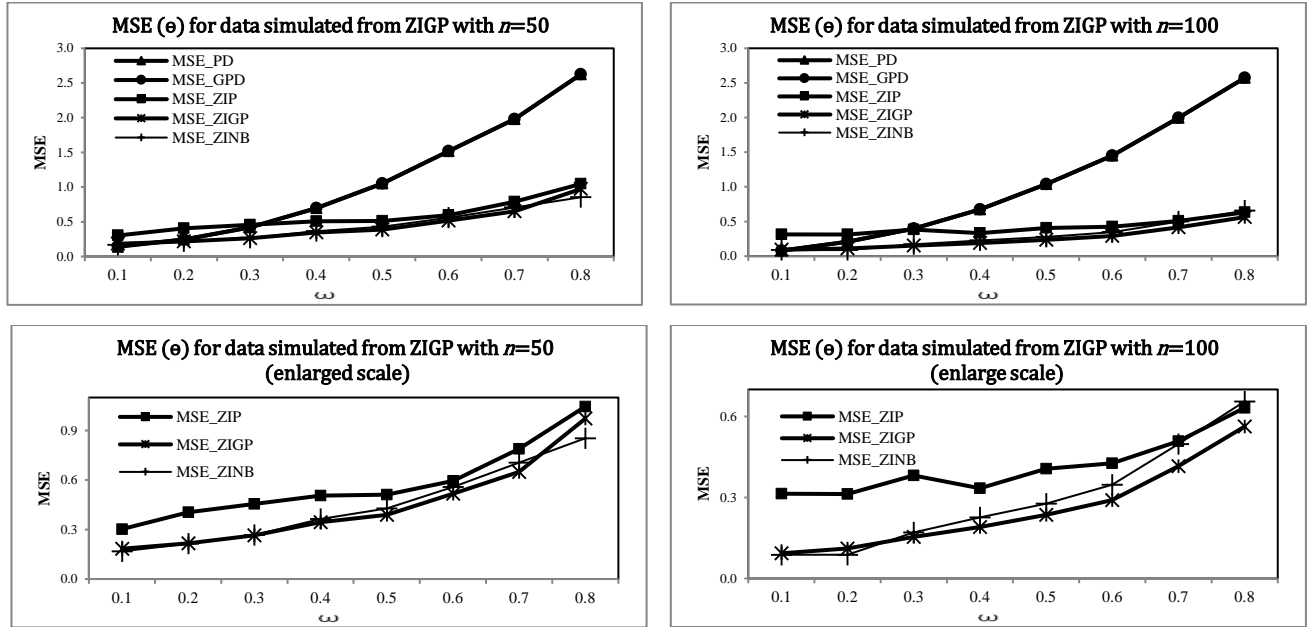
**Table 3.1: MSE, SE and bias for mean parameter  $\theta$  for  $ZIGP(2, 1.5, \omega_0)$**

$\omega_0$		$n = 50$					$n = 100$				
		P	GP	ZIP	ZIGP	ZINB	P	GP	ZIP	ZIGP	ZINB
0.1	MSE	0.1395	0.1395	0.3026	0.1836	0.1677	0.0797	0.0797	0.3136	0.0931	0.0880
	SE	0.3020	0.3020	0.4204	0.4192	0.4060	0.2086	0.2086	0.2848	0.3016	0.2968
	BIAS	-0.2200	-0.2200	0.4828	0.0905	0.0566	-0.1904	-0.1904	0.4823	0.0487	0.0081
0.2	MSE	0.2449	0.2449	0.4047	0.2160	0.2145	0.2067	0.2067	0.3127	0.1116	0.1168
	SE	0.2836	0.2836	0.4330	0.4639	0.4635	0.2058	0.2058	0.3059	0.3339	0.3398
	BIAS	-0.4058	-0.4058	0.4665	0.0360	-0.0121	-0.4055	-0.4055	0.4683	0.0195	-0.0401
0.3	MSE	0.4210	0.4210	0.4545	0.2641	0.2660	0.3984	0.3984	0.3814	0.1542	0.1700
	SE	0.2757	0.2757	0.4752	0.5092	0.5161	0.2053	0.2053	0.3501	0.3910	0.4112
	BIAS	-0.5874	-0.5874	0.4787	0.0731	0.0156	-0.5970	-0.5970	0.5090	0.0411	-0.0353
0.4	MSE	0.6977	0.6977	0.5058	0.3441	0.3637	0.6762	0.6762	0.3336	0.1904	0.2258
	SE	0.2824	0.2824	0.5235	0.5858	0.6030	0.1875	0.1875	0.3642	0.4361	0.4639
	BIAS	-0.7862	-0.7862	0.4820	0.0412	-0.0290	-0.8007	-0.8007	0.4486	-0.0236	-0.1050
0.5	MSE	1.0505	1.0505	0.5111	0.3888	0.4282	1.0375	1.0375	0.4069	0.2352	0.2767
	SE	0.2674	0.2674	0.5438	0.6232	0.6536	0.1890	0.1890	0.4206	0.4850	0.5221
	BIAS	-0.9895	-0.9895	0.4647	0.0345	-0.0437	-1.0009	-1.0009	0.4800	0.0210	-0.0681
0.6	MSE	1.5167	1.5167	0.5952	0.5171	0.5582	1.4512	1.4512	0.4266	0.2902	0.3466
	SE	0.2312	0.2312	0.6385	0.7190	0.7470	0.1629	0.1629	0.4267	0.5382	0.5862
	BIAS	-1.2097	-1.2097	0.4340	0.0349	-0.0368	-1.1936	-1.1936	0.4949	0.0337	-0.0602
0.7	MSE	1.9783	1.9783	0.7880	0.6492	0.7051	1.9944	1.9944	0.5084	0.4155	0.4975
	SE	0.2123	0.2123	0.7288	0.8010	0.8405	0.1448	0.1448	0.5401	0.6452	0.7003
	BIAS	-1.3904	-1.3904	0.5079	0.0944	0.0093	-1.4048	-1.4048	0.4661	0.0085	-0.0899
0.8	MSE	2.6249	2.6249	1.0453	0.9724	0.8532	2.5724	2.5724	0.6319	0.5626	0.6557
	SE	0.1644	0.1644	0.9396	0.9870	0.9859	0.1267	0.1267	0.6471	0.7507	0.8059
	BIAS	-1.6118	-1.6118	0.4052	0.0139	-0.0946	-1.5989	-1.5989	0.4626	0.0108	-0.0867

In terms of MSE, the MLE of  $\theta$  for Poisson and GP model performs much poorer than the ZIP, ZIGP and ZINB model irrespective of the values of zero-inflation parameter  $\omega$  (Fig.3.1-3.2). Fig. 3.3 and 3.4 depicts that ZIGP model has least MSE among the three zero-inflated models and hence ZIGP is the most suitable model to fit the data than ZIP and ZINB for both situations  $n = 50, 100$ . Similar results are observed for  $\theta = 5$  and  $n = 250$ .

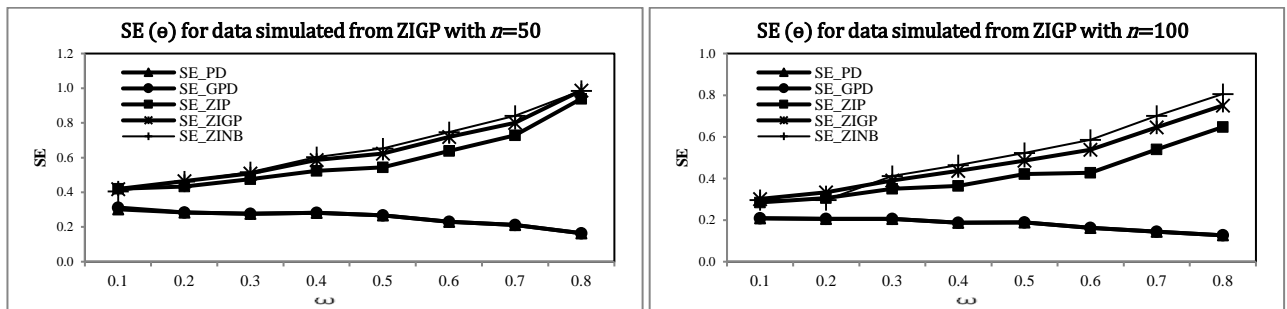


**Figure 3.1-3.4 MSE of estimated mean ( $\theta$ ) for ZIGP distribution**



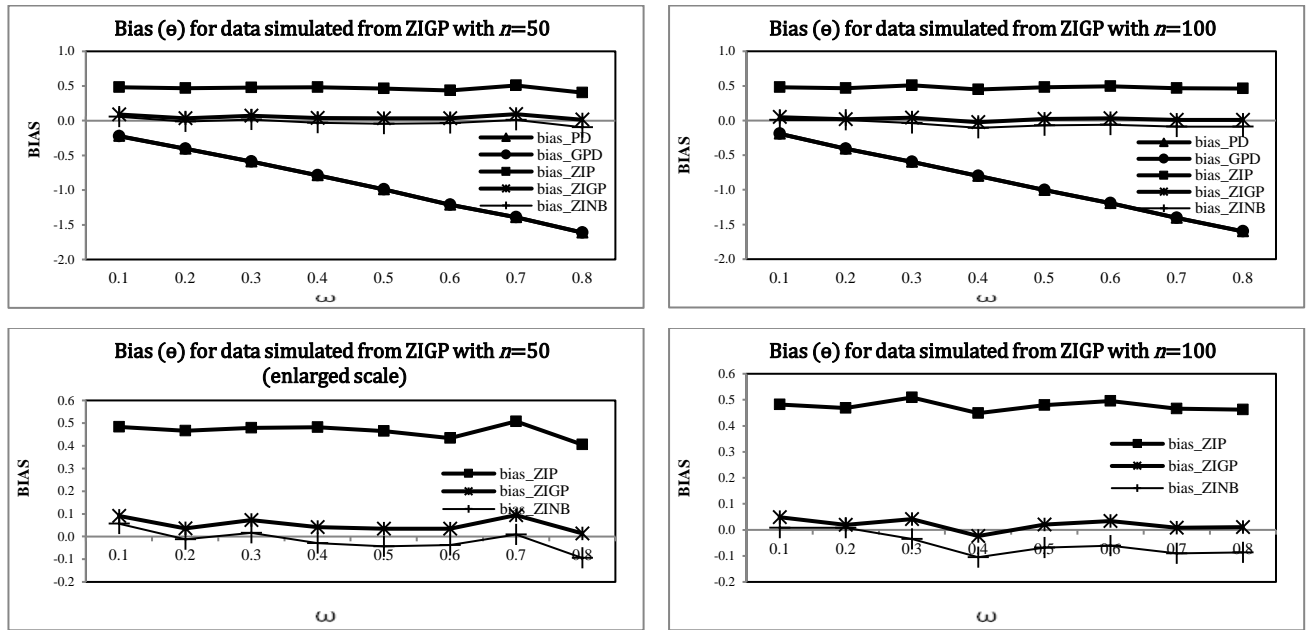
On the contrary, the SE of the estimate of the mean parameter in Fig. 3.5-3.6 shows that Poisson distribution (PD) and GPD have almost equal and lowest SE. Among zero-inflated models, ZIP model has least SE. The average SE for zero-inflated models increases as the value of inflation parameter  $\omega$  increases, regardless of the mean/sample size.

**Figure 3.5-3.6 SE of estimated mean of ZIGP distribution**



Further, Fig. 3.7-3.10 shows the behaviour of average bias estimates of the mean parameter. It can be seen that, PD and GPD underestimate the mean parameter while the ZIP model overestimates it. ZIGP and ZINB model have the least and almost near zero bias among all other models. The mean parameter estimates gradually became worse (negatively biased) for PD and GPD as the value of the inflation parameter  $\omega$  increases.

**Figure 3.7-3.10 Bias of estimated mean ( $\theta$ ) of ZIGP distribution**



Although, small SE means a smaller variability, this is not necessarily an advantage, especially if the estimator is biased as noted when the PD and GPD were fitted to a zero-inflated data.

Even though the PD and GPD both have the smallest SE in majority of the situations, both dramatically underestimate the mean. That is, the bias is driving the relationship in the statistical inference for both the PD and GPD. When the true distribution was ZIGP, the ZIP, ZIGP and ZINB models tracked very close to each other in case of all the sample sizes.

**The important observations from a simulation study are reported below.**

The average bias results point out that ignoring the zero-inflated nature of the data may result in a substantial underestimation of the mean and subsequently lead to insignificant findings. The bias of the mean parameter, by itself, does not allow definite conclusions about the statistical inference. Hence, one should not only look at the estimate of the mean parameter, but also at the inflation parameter. We explore the scenario through the box-plot of estimated mean and inflation parameter for 500 random samples each of size  $n = 50$  from ZIGP distribution with  $\theta = 2, \varphi = 1.5$  and  $\omega = 0.5$  as in Fig. 3.11 and 3.12 respectively.

**Figure 3.11-3.14** Boxplot of  $\hat{\theta}^d$  and  $\hat{\omega}^d$  of 500 random samples from  $ZIGP(2, 1.5, 0.5)$  distribution.

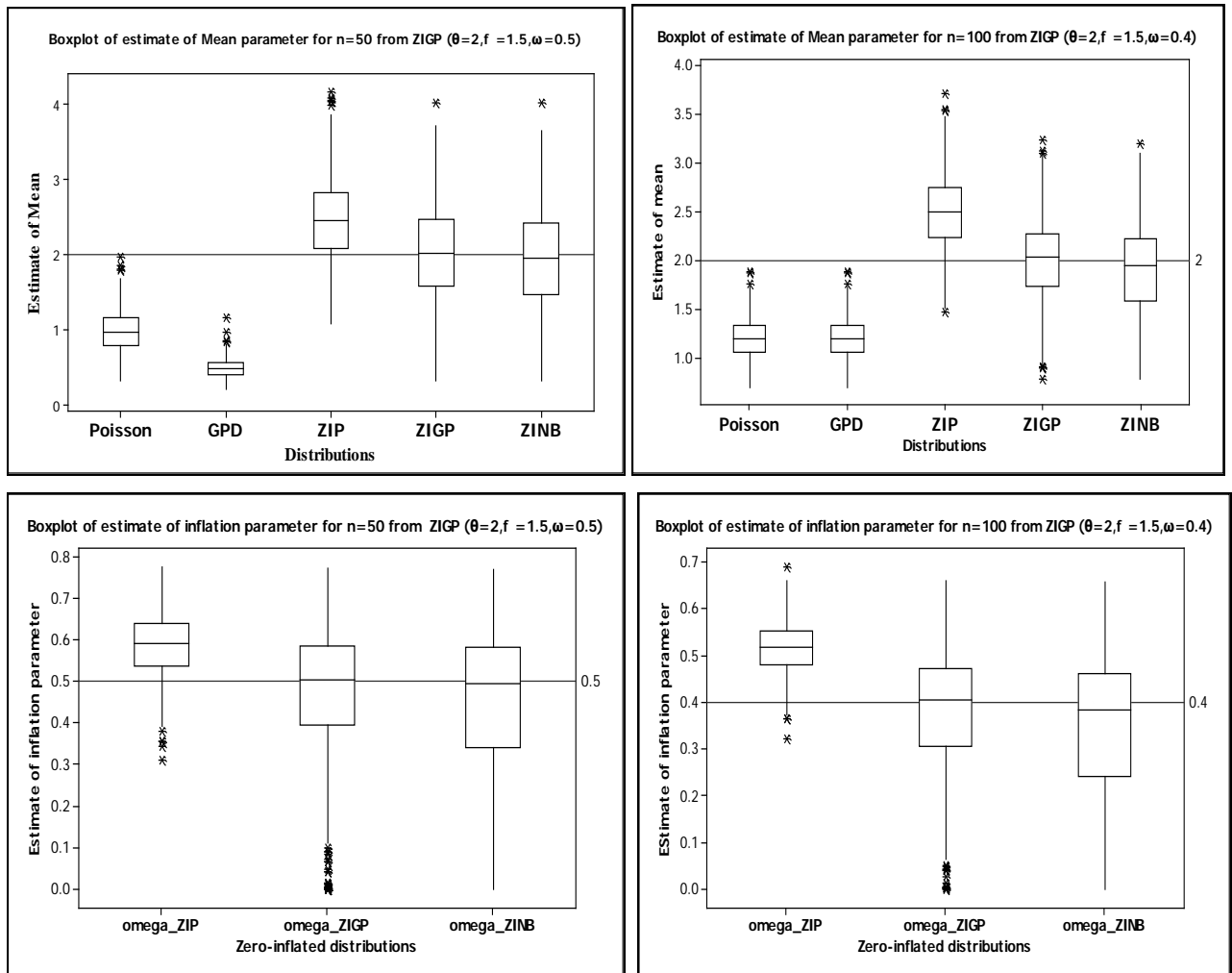


Fig.3.11, 3.12 shows box-plot of estimated mean of 500 random samples of size  $n = 50, 100$  from  $ZIGP(2,1.5,0.5)$  and  $ZIGP(2,1.5,0.4)$  respectively. It can be seen that, Poisson and GP model underestimate whereas ZIP model overestimates the mean. While ZIGP and ZINB model correctly identify the mean. Fig. 3.13 and 3.14 shows box-plot of MLE of inflation parameter. The box-plot of the MLE of inflation parameter shows that ZIGP model correctly identify the inflation parameter  $\omega$  with lots of outliers to the lower side. While ZINB model underestimate and the ZIP model overestimate the inflation parameter  $\omega$ .

There are different measures of detecting overdispersion and zero-inflation for the zero-inflated count data. One such measure is discussed in the following Section.

#### 4. Zero-Inflation Index

Puig and Valero (2006) introduced zero-inflation index (ZI index) and dispersion index as a measure of detecting zero-inflation and overdispersion respectively from the Poisson distribution. If  $Y$  is a nonnegative integer random variable (count variable) with mean  $\theta$  and  $p_0$  is the proportion of zeros in a random sample of size  $n$  then ZI index  $z_i$  for the sample is defined as,

$$z_i = 1 + \log(p_0)/\theta$$

For the large sample from Poisson distribution  $z_i$  is close to zero with high probability. We perform a simulation work to study the behaviour of  $z_i$  for different sample sizes across different values of zero-inflation parameter.

**Table 4.1 Average zero-inflation index for samples of size  $n = 15, 25, 50, 100$  from ZIGP distribution across inflation parameter  $\omega$**

Sample size ( $n$ )	Average zero-inflation index ( $z_i$ )							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
15	0.414470	0.512770	0.611025	0.696613	0.761046	0.818421	0.864039	0.908972
25	0.418849	0.536377	0.629724	0.703338	0.742960	0.815451	0.872185	0.919649
50	0.457601	0.548460	0.627525	0.700465	0.764217	0.823715	0.873006	0.917429
100	0.445320	0.550769	0.638102	0.696761	0.768757	0.823649	0.874969	0.919606

Table 4.1 shows the ZI index for 500 random samples of size  $n = 15, 25, 50, 100$  from ZIGP distribution with  $\theta = 2, \varphi = 1.5$  across zero-inflation parameter  $\omega_0 = 0.1, 0.2, \dots, 0.8$ . It can be observed that ZI index is positive and always overestimates the true parameter for all the sample sizes.

**Figure 4.1 Box plot of ZI index for random samples from  $ZIGP(\theta, \varphi, \omega) |_{\theta=2, \varphi=1.5, \omega=\omega_0}$ , across zero inflation parameter with  $n = 50, 100$ .**

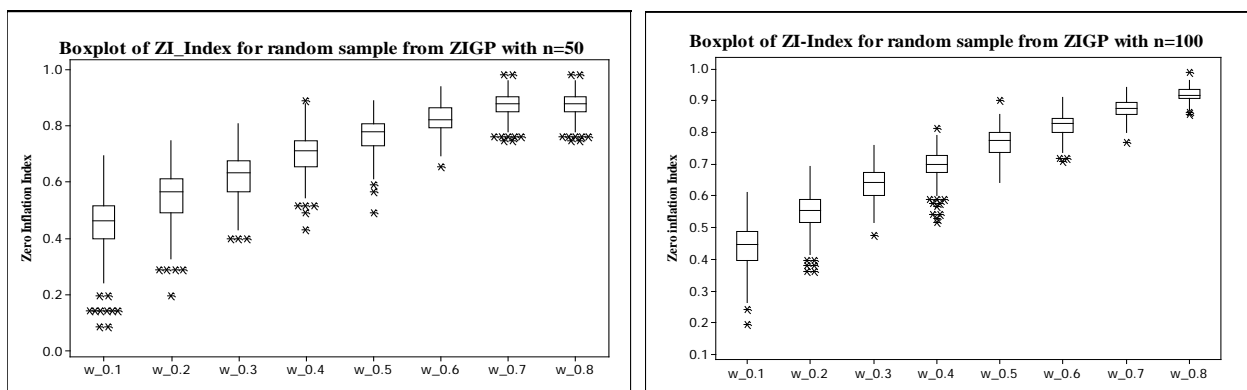


Fig. 4.1 shows that ZI index is positive almost for all samples. Even though ZI index is not defined as an estimate of inflation parameter, one can use it to get a rough idea about zero-inflation parameter  $\omega$ . Further, ZI index would always overestimate the true parameter value of  $\omega_0$  if treated as an estimate of  $\omega$  while for higher values of  $\omega$  it becomes more and more consistent. It can also be seen that this behaviour of ZI index is independent of sample size.

## 5. Parameter Estimation in ZIGP Distribution

In this section, we briefly review the methods of parameter estimation concerning ZIGP distribution and propose estimators for parameters based on the ratio of first two moments and proportion of zeros of ZIGP distribution which perform as good as MLE and even more efficient than the MLE.

We consider the zero-inflated GP model introduced by Consul and Famoye (1992) and denote it as  $ZIGP(\theta, \varphi, \omega)$ . The pmf of  $Y \sim ZIGP(\theta, \varphi, \omega)$  is,

$$P(Y = y | \theta, \varphi, \omega) = \begin{cases} \omega + (1 - \omega) e^{-\theta/\varphi} & y = 0 \\ (1 - \omega) \theta (\theta + y(\varphi - 1))^{y-1} \frac{\varphi^{-y}}{y!} \exp\left(-\frac{(\theta + y(\varphi - 1))}{\varphi}\right) & y > 0 \end{cases}$$

where  $0 \leq \omega \leq 1$ ,  $\theta > 0$ ,  $\varphi \geq \max\left\{\frac{1}{2}, 1 - \frac{\theta}{m}\right\}$  and  $m(m > 4)$  is a largest natural number for which  $\theta + m(\varphi - 1) > 0$  when  $\varphi < 1$ .

We discuss methods of estimation of parameters of ZIGP distribution.

### 5.1 Maximum Likelihood Estimation

The log-likelihood function based on a random sample  $y_1, y_2, \dots, y_n$  from  $ZIGP(\theta, \varphi, \omega)$  distribution is,

$$\begin{aligned} \log L(\theta, \varphi, \omega | y_1, y_2, \dots, y_n) &= \sum_{i=1}^n I_{(y_i=0)} \log\left(\omega + (1 - \omega) e^{-\frac{\theta}{\varphi}}\right) + \sum_{i=1}^n I_{(y_i>0)} \left\{ \frac{-(\theta + y_i(\varphi - 1))}{\varphi} \right\} \\ &+ \sum_{i=1}^n I_{(y_i>0)} \{ \log(1 - \omega) \mp (y_i - 1) \log(\theta + y_i(\varphi - 1)) - y_i \log \varphi - \log y_i! \} \end{aligned}$$

Iterative methods are used to obtain MLE of the parameters of ZIGP model. Czado and Min (2005) proved the consistency and asymptotic normality of a solution of maximum likelihood estimators

(MLE) in ZIGP regression model. We use the ZIGP package in R software to get the MLE of ZIGP model.

## 5.2 Estimator based on first two moments and proportion of zeros (MOZE)

Wagh and Kamalja (2017b) compared the estimates of GPD parameter in terms of bias, MSE, covariance using the method of moment estimation and MLE through a simulation study and demonstrated that ME performs better or equally good as compared to MLE when the sample size is small. Wagh and Kamalja (2017c) proposed an estimator referred to as probability estimator of inflation parameter of ZIP distribution based on moment estimator of the mean parameter and compare its performance through a simulation study. Famoye (1997) estimates the parameters of GNBD by the method of maximum likelihood, MOZE, first two moments and the ratio of the first two frequencies (MORA) and Minimum Chi-square (MC) and compares them through asymptotic relative efficiency. In this section, we introduce new estimators of parameters of ZIGP distribution which are based on first two sample moments and proportion of zeros; but the method of obtaining the estimates is somewhat different than that of Famoye (1997). We describe this method in the following.

The first two population moments of  $ZIGP(\theta, \varphi, \omega)$  distribution are,

$$E(Y) = (1 - \omega)\theta, \quad E(Y^2) = (1 - \omega)\theta(\theta + \varphi^2)$$

Let  $M_1 = \frac{1}{n} \sum_{i=1}^n Y_i$  and  $M_2 = \frac{1}{n} \sum_{i=1}^n Y_i^2$  be the first and second sample moment respectively, based on a random sample  $y_1, y_2, \dots, y_n$  from  $ZIGP(\theta, \varphi, \omega)$  distribution. The moment equation obtained by equating first sample moment with the corresponding population moment of ZIGP distribution when simplified for  $\theta$  gives,

$$\theta = \frac{M_1}{1-\omega} \tag{5.1}$$

Equating ratio of  $M_2$  and  $M_1$  with the ratio of corresponding population moments of ZIGP distribution gives,

$$\frac{M_2}{M_1} = \theta + \varphi^2$$

Simplifying the above equation for  $\varphi$ , we get,

$$\varphi = \sqrt{\frac{M_2}{M_1} - \frac{M_1}{1-\omega}} \quad (5.2)$$

Let  $p_0 = P(Y = 0)$  and  $\hat{p}_0 = \frac{n_0}{n}$  be the corresponding empirical probability where  $n_0$  is the frequency of zero in a random sample. Then

$$p_0 = \omega + (1 - \omega)e^{-\theta/\varphi}$$

Simplifying the expression  $\hat{p}_0 = p_0$  for  $\omega$ , we get the following probability equation.

$$\omega = \frac{\hat{p}_0 - e^{-\theta/\varphi}}{1 - e^{-\theta/\varphi}} \quad (5.3)$$

Thus  $\omega$  is in terms of  $\theta$  and  $\varphi$ . An estimate of  $\omega$  based on (5.3) using ME of  $\theta$  and  $\varphi$  is calculated. We use the following steps to obtain estimates of parameters.

i) Obtain the initial estimates  $\hat{\theta}_0$  and  $\hat{\varphi}_0$  of  $\theta$  and  $\varphi$  with  $\hat{\omega}_0 = \hat{p}_0$ . That is

$$\hat{\theta}_0 = \frac{M_1}{1-\hat{\omega}_0} \quad \text{and} \quad \hat{\varphi}_0 = \sqrt{\frac{M_2}{M_1} - \frac{M_1}{1-\hat{\omega}_0}}$$

ii) The improved estimates  $\hat{\omega}_3, \hat{\theta}_3$  and  $\hat{\varphi}_3$  are obtained iteratively using (5.3), (5.1) and (5.2) respectively as follows for  $i = 1, 2, 3$ .

$$\begin{aligned} \hat{\omega}_i &= \frac{\hat{p}_0 - e^{-(\hat{\theta}_{i-1}/\hat{\varphi}_{i-1})}}{1 - e^{-(\hat{\theta}_{i-1}/\hat{\varphi}_{i-1})}} \\ \hat{\theta}_i &= \frac{M_1}{1-\hat{\omega}_i} \\ \hat{\varphi}_i &= \sqrt{\frac{M_2}{M_1} - \frac{M_1}{1-\hat{\omega}_i}} \end{aligned} \quad (5.4)$$

We use only three iterations since at this stage we get an estimate, which is more or almost equally efficient than corresponding MLEs. In the next section, we demonstrate the efficiency of the proposed estimators with respect to MLE based on simulation.

## 6. Comparison of estimate of parameters of ZIGP distribution

We perform a simulation study to compare the performance of proposed estimates of parameters of ZIGP distribution through MSE, bias, and SE. The simulation study in this paper

demonstrates that the new estimators  $\hat{\theta}_3, \hat{\varphi}_3$  of  $\theta$  and  $\varphi$  are almost equally efficient as that of respective MLEs while  $\hat{\omega}_3$  is much more efficient than  $\hat{\omega}_{MLE}$ . We discuss the steps of simulation study systematically in the following.

Generate  $m$  random samples each of size  $n$  from ZIGP distribution and obtain MLE  $\hat{\theta}_{MLE}, \hat{\varphi}_{MLE}, \hat{\omega}_{MLE}$ , of  $\theta, \varphi$  and  $\omega$  using ZIGP package in R-Software as discussed by Erhardt (2011). Calculate the proposed estimators  $\hat{\omega}_3, \hat{\theta}_3$  and  $\hat{\varphi}_3$  for each sample. The algorithm is as follows.

- i)  $m = 500$  random samples each of size  $n$  are generated from  $ZIGP(\theta, \varphi, \omega)$  for  $\theta = 2, \varphi = 1.5$  and  $\omega = \omega_0$  where  $\omega_0 \in (0,1)$ .
- ii) Compute  $\hat{\theta}_{MLE}^{(i)}, \hat{\varphi}_{MLE}^{(i)}, \hat{\omega}_{MLE}^{(i)}$  and  $\hat{\theta}_3^{(i)}, \hat{\varphi}_3^{(i)}, \hat{\omega}_3^{(i)}$  for each of the  $i^{th}$  sample.
- iii) Evaluate  $MSE(\hat{\theta}_{MLE})$  and  $MSE(\hat{\theta}_3)$  using following formulae.

$$MSE(\hat{\theta}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\theta}_{MLE}^{(i)} - \theta)^2}{m}, \quad MSE(\hat{\theta}_3) = \frac{\sum_{i=1}^m (\hat{\theta}_3^{(i)} - \theta)^2}{m}.$$

Similarly, evaluate  $MSE(\hat{\varphi}_{MLE}), MSE(\hat{\varphi}_3), MSE(\hat{\omega}_{MLE})$  and  $MSE(\hat{\omega}_3)$  by using following formulae.

$$\begin{aligned} MSE(\hat{\varphi}_{MLE}) &= \frac{\sum_{i=1}^m (\hat{\varphi}_{MLE}^{(i)} - \varphi)^2}{m}, & MSE(\hat{\varphi}_3) &= \frac{\sum_{i=1}^m (\hat{\varphi}_3^{(i)} - \varphi)^2}{m} \\ MSE(\hat{\omega}_{MLE}) &= \frac{\sum_{i=1}^m (\hat{\omega}_{MLE}^{(i)} - \omega)^2}{m}, & MSE(\hat{\omega}_3) &= \frac{\sum_{i=1}^m (\hat{\omega}_3^{(i)} - \omega)^2}{m} \end{aligned}$$

- iv) Further evaluate the relative efficiency of all the proposed estimates with respect to respective MLE as follows.

$$RE(\hat{\theta}_3 | \hat{\theta}_{MLE}) = \frac{MSE(\hat{\theta}_{MLE})}{MSE(\hat{\theta}_3)}, \quad RE(\hat{\varphi}_3 | \hat{\varphi}_{MLE}) = \frac{MSE(\hat{\varphi}_{MLE})}{MSE(\hat{\varphi}_3)}$$

$$RE(\hat{\omega}_3 | \hat{\omega}_{MLE}) = \frac{MSE(\hat{\omega}_{MLE})}{MSE(\hat{\omega}_3)}$$

- v) The variance of the proposed estimates and MLEs is calculated as follows.

$$var(\hat{\theta}_3) = \frac{\sum_{i=1}^m (\hat{\theta}_3^{(i)} - \bar{\theta}_3)^2}{m-1}, \quad var(\hat{\theta}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\theta}_{MLE}^{(i)} - \bar{\theta}_{MLE})^2}{m-1},$$

$$var(\hat{\varphi}_3) = \frac{\sum_{i=1}^m (\hat{\varphi}_3^{(i)} - \bar{\varphi}_3)^2}{m-1}, \quad var(\hat{\varphi}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\varphi}_{MLE}^{(i)} - \bar{\varphi}_{MLE})^2}{m-1},$$

$$var(\hat{\omega}_3) = \frac{\sum_{i=1}^m (\hat{\omega}_3^{(i)} - \bar{\omega}_3)^2}{m-1}, \quad var(\hat{\omega}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\omega}_{MLE}^{(i)} - \bar{\omega}_{MLE})^2}{m-1}$$



$$\text{where } \bar{\hat{\theta}}_3 = \frac{\sum_{i=1}^m \hat{\theta}_3^{(i)}}{m}, \quad \bar{\hat{\theta}}_{MLE} = \frac{\sum_{i=1}^m \hat{\theta}_{MLE}^{(i)}}{m}, \quad \bar{\hat{\phi}}_3 = \frac{\sum_{i=1}^m \hat{\phi}_3^{(i)}}{m},$$

$$\bar{\hat{\phi}}_{MLE} = \frac{\sum_{i=1}^m \hat{\phi}_{MLE}^{(i)}}{m}, \quad \bar{\hat{\omega}}_3 = \frac{\sum_{i=1}^m \hat{\omega}_3^{(i)}}{m}, \quad \bar{\hat{\omega}}_{MLE} = \frac{\sum_{i=1}^m \hat{\omega}_{MLE}^{(i)}}{m}.$$

We evaluate standard errors  $SE(\hat{\theta}_3)$ ,  $SE(\hat{\theta}_{MLE})$ ,  $SE(\hat{\phi}_3)$ ,  $SE(\hat{\phi}_{MLE})$ ,  $SE(\hat{\omega}_3)$  and  $SE(\hat{\omega}_{MLE})$  of the estimators from their respective variances.

vi) We also calculate average biases in the respective estimators as follows.

$$\text{bias}(\hat{\theta}_3) = \frac{\sum_{i=1}^m (\hat{\theta}_3^{(i)} - \theta)}{m}, \quad \text{bias}(\hat{\theta}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\theta}_{MLE}^{(i)} - \theta)}{m},$$

$$\text{bias}(\hat{\phi}_3) = \frac{\sum_{i=1}^m (\hat{\phi}_3^{(i)} - \phi)}{m}, \quad \text{bias}(\hat{\phi}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\phi}_{MLE}^{(i)} - \phi)}{m},$$

$$\text{bias}(\hat{\omega}_3) = \frac{\sum_{i=1}^m (\hat{\omega}_3^{(i)} - \omega)}{m}, \quad \text{bias}(\hat{\omega}_{MLE}) = \frac{\sum_{i=1}^m (\hat{\omega}_{MLE}^{(i)} - \omega)}{m}.$$

Steps (i) to (vi) are repeated for  $\omega_0 = 0.1, 0.2, \dots, 0.8$ .

The simulation study is performed for the sample sizes  $n = 50, 100, 200, 250$ . Table 6.1 and 6.2 shows the results for  $n = 50, 100$  (the results corresponding to  $n = 200, 250$  are not presented in the paper due to the almost similar behaviour).

**Table 6.1 MSE, SE, Bias and Relative Efficiency for  $\theta, \phi, \omega$  of ZIGP distribution for  $n = 50$ .**

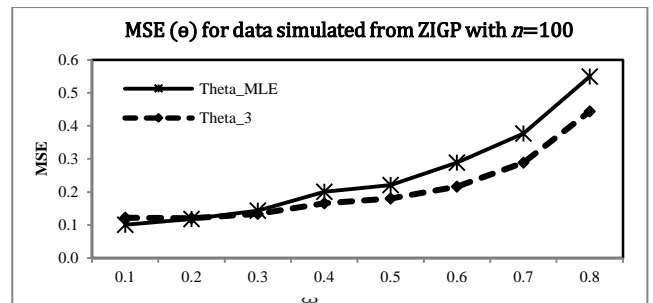
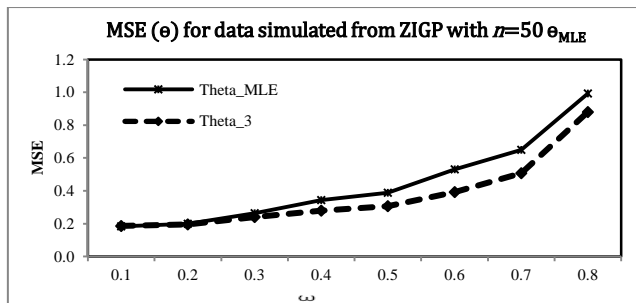
		Parameter $\theta$				Parameter $\phi$				Parameter $\omega$			
$\omega_0$	Estimate type	MSE	SE	BIAS	RE( $\hat{\theta}$   $\hat{\theta}_{MLE}$ )	SE	BIAS	RE( $\hat{\phi}$   $\hat{\phi}_{MLE}$ )	MSE	SE	BIAS	RE( $\hat{\omega}$   $\hat{\omega}_{MLE}$ )	
0.1	MLE	0.1836	0.4192	0.0905		0.0668	0.2466	-0.0783		0.0150	0.1165	0.0379	
	MOZE	0.1863	0.3760	0.2128	0.9851	0.0733	0.2487	-0.1079	0.9111	0.0197	0.1090	0.0886	0.7609
0.2	MLE	0.2000	0.4432	0.0630		0.0833	0.2830	-0.0577		0.0191	0.1381	-0.0035	
	MOZE	0.1955	0.3971	0.1954	1.0229	0.0850	0.2693	-0.1121	0.9800	0.0158	0.1074	0.0658	1.2027
0.3	MLE	0.2641	0.5092	0.0731		0.0847	0.2792	-0.0828		0.0232	0.1524	-0.0060	
	MOZE	0.2393	0.4307	0.2328	1.1032	0.0851	0.2552	-0.1420	0.9941	0.0158	0.1114	0.0584	1.4702
0.4	MLE	0.3441	0.5858	0.0412		0.1023	0.3142	-0.0615		0.0290	0.1684	-0.0258	
	MOZE	0.2788	0.4775	0.2264	1.2344	0.1034	0.2934	-0.1321	0.9894	0.0150	0.1136	0.0465	1.9261
0.5	MLE	0.3888	0.6232	0.0345		0.1343	0.3583	-0.0784		0.0316	0.1753	-0.0314	
	MOZE	0.3054	0.4996	0.2374	1.2732	0.1257	0.3121	-0.1687	1.0683	0.0124	0.1033	0.0417	2.5568
0.6	MLE	0.5308	0.7293	0.0025		0.1614	0.3965	-0.0673		0.0447	0.1812	-0.0135	
	MOZE	0.3913	0.5816	0.2317	1.3564	0.1569	0.3586	-0.1692	1.0283	0.0114	0.1006	0.0353	3.9383
0.7	MLE	0.6492	0.8010	0.0944		0.2010	0.4374	-0.1001		0.0459	0.1751	-0.0494	
	MOZE	0.5068	0.6371	0.3191	1.2810	0.1654	0.3655	-0.1790	1.2151	0.0070	0.0795	0.0268	6.5322
0.8	MLE	0.9919	0.9959	0.0446		0.2603	0.4956	-0.1231		0.0517	0.2195	-0.0603	
	MOZE	0.8803	0.7878	0.5108	1.1268	0.2605	0.4321	-0.2722	0.9991	0.0051	0.0684	0.0209	10.1263

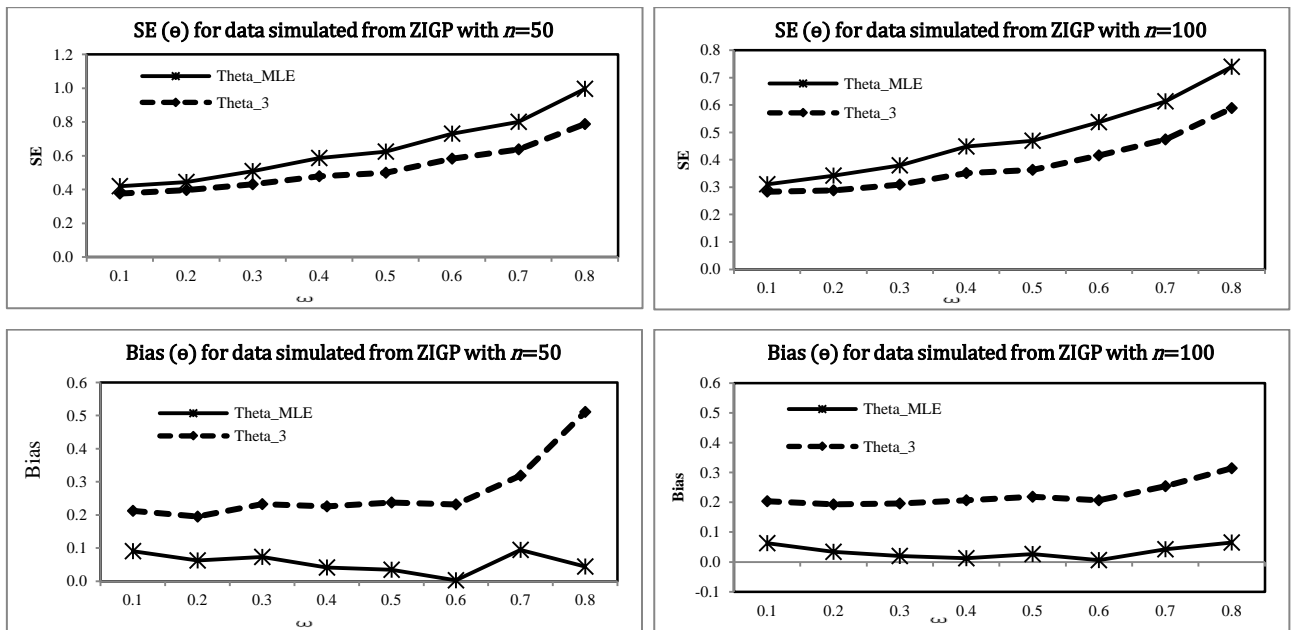
**Table 6.2 MSE, SE, Bias and Relative Efficiency for  $\theta, \phi, \omega$  of ZIGP distribution for  $n = 100$ .**

		Parameter $\theta$				Parameter $\varphi$				Parameter $\omega$			
$\omega_0$	Estimate type	MSE	SE	BIAS	$RE(\hat{\theta} \hat{\theta}_{MLE})$	MSE	SE	BIAS	$RE(\hat{\varphi} \hat{\varphi}_{MLE})$	MSE	SE	BIAS	$RE(\hat{\omega} \hat{\omega}_{MLE})$
0.1	MLE	0.1003	0.3108	0.0629		0.0321	0.1750	-0.0385		0.0091	0.0942	0.0160	
	MOZE	0.1217	0.2832	0.2041	0.8244	0.0373	0.1739	-0.0845	0.8589	0.0120	0.0787	0.0763	0.7594
0.2	MLE	0.1177	0.3417	0.0339		0.0401	0.1979	-0.0322		0.0114	0.1067	0.0014	
	MOZE	0.1204	0.2885	0.1934	0.9772	0.0444	0.1931	-0.0849	0.9035	0.0108	0.0807	0.0659	1.0493
0.3	MLE	0.1438	0.3791	0.0196		0.0475	0.2171	-0.0211		0.0150	0.1221	-	
	MOZE	0.1337	0.3091	0.1959	1.0755	0.0462	0.1985	-0.0827	1.0288	0.0098	0.0821	0.0556	1.5277
0.4	MLE	0.2008	0.4483	0.0129		0.0584	0.2415	-0.0144		0.0181	0.1331	-	
	MOZE	0.1658	0.3509	0.2073	1.2108	0.0561	0.2197	-0.0888	1.0421	0.0084	0.0780	0.0477	2.1712
0.5	MLE	0.2206	0.4694	0.0259		0.0733	0.2683	-0.0377		0.0156	0.1243	-	
	MOZE	0.1797	0.3632	0.2191	1.2277	0.0673	0.2332	-0.1141	1.0884	0.0071	0.0714	0.0443	2.2139
0.6	MLE	0.2882	0.5374	0.0058		0.0906	0.2994	-0.0346		0.0179	0.1319	-	
	MOZE	0.2155	0.4160	0.2069	1.3371	0.0816	0.2639	-0.1100	1.1103	0.0058	0.0685	0.0328	3.1108
0.7	MLE	0.3769	0.6131	0.0421		0.1223	0.3454	-0.0566		0.0187	0.1344	-	
	MOZE	0.2890	0.4741	0.2543	1.3040	0.1095	0.2960	-0.1484	1.1169	0.0046	0.0639	0.0235	4.0450
0.8	MLE	0.5492	0.7390	0.0646		0.1792	0.4223	-0.0348		0.0178	0.1305	-	
	MOZE	0.4438	0.5879	0.3143	1.2376	0.1626	0.3706	-0.1597	1.1018	0.0027	0.0490	0.0188	6.4612

Table 6.1 and 6.2 shows that MOZE and MLE of ZIGP distribution performs equally for the parameters  $\theta$ ,  $\varphi$  and  $\omega$ . We observed, through the relative efficiency of MOZE with respect to MLE for all the parameters of ZIGP distribution that, MOZE performs as good as the MLE for the sample sizes  $n = 50, 100$ . The results can be visualized through the Fig. 6.1-6.6 for  $n = 50, 100$  which clearly shows the behaviour of the MLE and the MOZE for all the parameters across the values of  $\omega$ .

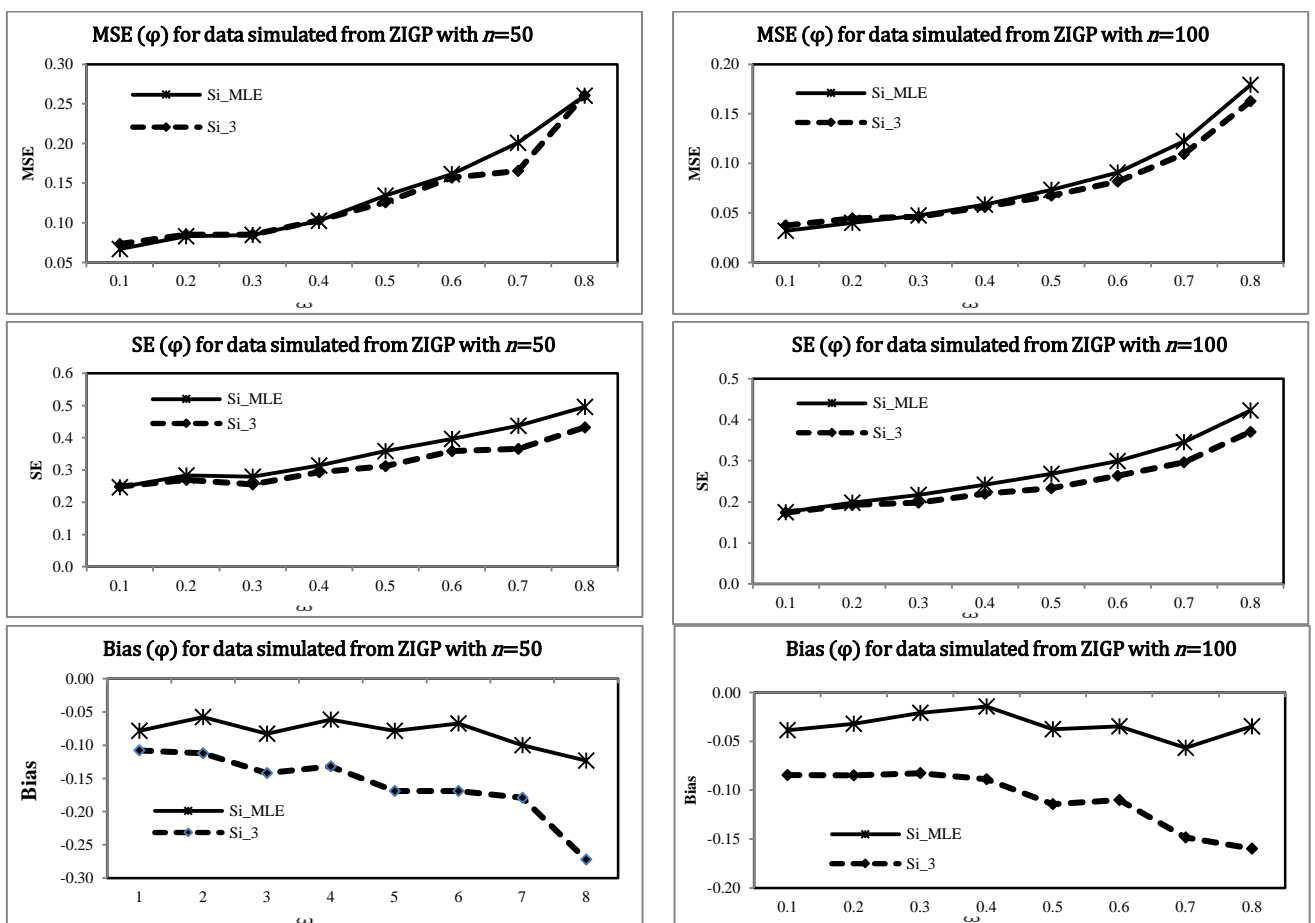
**Figure 6.1-6.6 MSE, SE and Bias of estimated Mean parameter  $\theta$  for ZIGP distribution for  $n = 50, 100$ .**





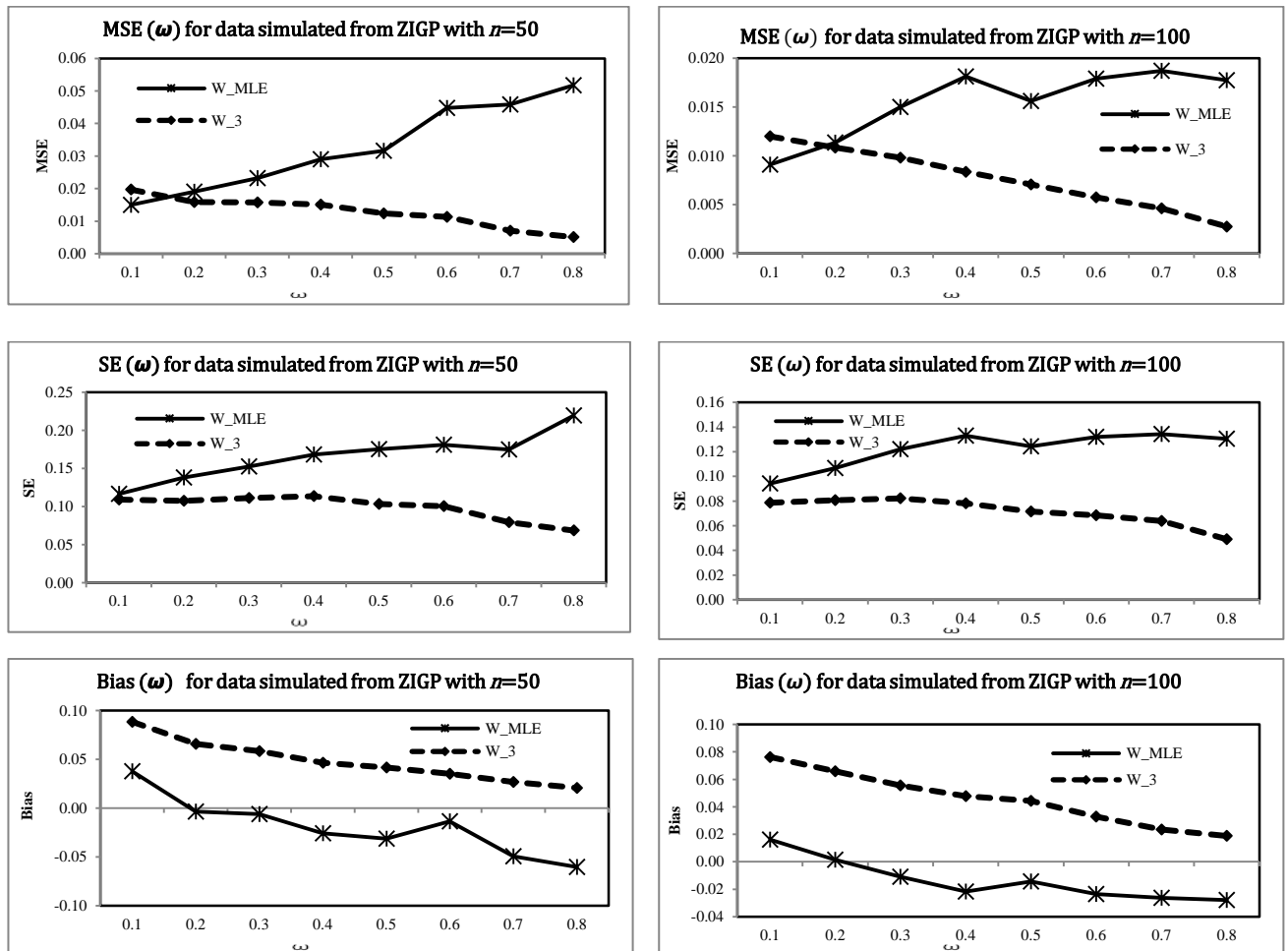
It can be observed that the MSE and SE of the mean parameter of  $\hat{\theta}_3$  is less than that of  $\hat{\theta}_{MLE}$ . Overall the MOZE of mean parameter  $\theta$  is equally efficient as MLE for ZIGP distribution.

**Figure 6.7-6.12** MSE, SE and Bias of estimated dispersion parameter  $\varphi$  for ZIGP distribution for  $n = 50, 100$ .



It can be observed that the MSE and SE for the dispersion parameter of  $\hat{\varphi}_3$  is less than that of  $\hat{\varphi}_{MLE}$ . Overall the MOZE of dispersion parameter  $\varphi$  is equally efficient as MLE for ZIGP distribution.

**Figure 6.13-6.18 MSE, SE and Bias of estimated inflation parameter  $\omega$  for ZIGP distribution for  $n = 50, 100$ .**



Further, we note that the MSE for inflation parameter  $\hat{\omega}_3$  is less than MSE of  $\hat{\omega}_{MLE}$ . Overall the MOZE of inflation parameter  $\omega$  is more efficient than MLE for ZIGP distribution. The boxplot (not included) shows that  $\hat{\omega}_3$  is more consistent than that of  $\hat{\omega}_{MLE}$ .

## 7. Fitting different count data models

For the purpose of demonstration, we fit Poisson, GPD, ZIP, ZIGP, and ZINB to various real life data sets and compare the values of chi-square statistic for goodness of fit and p-value at 5% level of significance. The results are reported in the following Table 7.1.

**Table 7.1 Fitting count data model to zero-inflated data sets**

Reference of the sample data set used	Details of the Sample data set	Sample Size	Parameter	Goodness of fit under different model and details of model fitting				
				Poisson	GPD	ZIP	ZIGP	ZINB
Santolino and Boucher (2009)	Spanish insurance claim data set	180	$\hat{\theta}$	4.6111	4.6109	7.4732	6.8259	6.4818
			$\hat{\phi}$	-	4.0728	-	2.7972	0.9968
			$\hat{\omega}$	-	-	0.3830	0.3245	0.2886
			$\chi^2$	1631634.44	33.4889	10776.9	16.9866	16.7737
			p-value	0.0000	0.0004	0.0000	0.0747	0.0795
Lawal (2012)	Injury counts for cleaner pre-WRATS	342	$\hat{\theta}$	0.3275	0.3275	0.8638	0.3282	0.3276
			$\hat{\phi}$	-	1.3289	-	1.3282	2.2257
			$\hat{\omega}$	-	-	0.6209	0.0025	0.0003
			$\chi^2$	82.4239	0.3461	3.4604	0.3445	0.3883
			p-value	0.0000	0.8411	0.1772	0.5572	0.5332
Lawal (2012)	Mosquito count from 492 houses in Western Kenya	492	$\hat{\theta}$	1.4858	1.4858	3.4341	2.1005	1.4871
			$\hat{\phi}$	-	2.6637	-	2.3434	3.2517
			$\hat{\omega}$	-	-	0.5673	0.2927	0.0009
			$\chi^2$	5.118E+10	15.0649	1399980	13.9254	15.7383
			p-value	0.0000	0.0579	0.0000	0.0525	0.0276
Rideout, Hinde, Demetrio (2001)	Data for 270 micro-propagated roots	270	$\hat{\theta}$	5.0593	5.0591	6.6222	6.5924	6.5889
			$\hat{\phi}$	-	2.2921	-	1.2812	0.1003
			$\hat{\omega}$	-	-	0.2360	0.2326	0.2322
			$\chi^2$	2462.3554	127.6170	61.1898	16.6173	15.6528
			p-value	0.0000	0.0000	0.0000	0.0833	0.1100
Simulated data from ZIGP (2, 1.5, 0.4) $x = 0, 1, 2, 3, 4, 5, 6, 8, 9$ $f = 55, 15, 11, 8, 2, 6, 1, 1,$		100	$\hat{\theta}$	1.2200	1.2199	2.4853	2.0869	2.0309
			$\hat{\phi}$	-	1.9087	-	1.4210	0.5246
			$\hat{\omega}$	-	-	0.5091	0.4154	0.6647
			$\chi^2$	104.4887	3.5509	4.7384	0.1201	0.0992
			p-value	0.0000	0.3142	0.1920	0.9417	0.9516
Simulated data from ZIGP (3, 1.3, 0.5) $x = 0: 8, 10, 11, 13, 14$ $f = 160, 22, 28, 26, 22, 19, 8, 5, 6, 1, 1, 1, 1$		300	$\hat{\theta}$	1.7267	1.7266	3.5988	3.4241	3.4119
			$\hat{\phi}$	-	2.3666	-	1.3190	0.2173
			$\hat{\omega}$	-	-	0.5202	0.4957	0.4939
			$\chi^2$	1286.3096	46.8497	21.2283	3.3325	3.3952
			p-value	0.0000	0.0000	0.0017	0.6489	0.6393

### 8. Conclusions

In this paper, we take a brief overview of different zero-inflated distributions that are popular for modelling count data having zero-inflation and overdispersion. We present a simulation study for choosing between zero-inflated models by comparing the performance of the estimates of Poisson, GPD, ZIP, ZIGP and ZINB model when the data come from ZIGP distribution. If the zero-inflated nature of the data is ignored and either a Poisson or GPD model is used to model a ZIGP data, we fail

to notice some significant relationships. To correctly identify the zero-inflated model one should consider an estimate of inflation parameter along with the estimate of the mean parameter.

We discuss the estimation of parameters of ZIGP distribution and propose estimators of parameters of ZIGP distribution based on the first two sample moments and proportion of zeros referred as MOZE and compares its performance with MLE through a simulation study. Through simulation study, it is observed that MOZE are almost equal or even sufficient than that of MLE of the parameters of ZIGP distribution.

## References

- [1] Cameron, A. and Trivedi, P. (1986). *Regression analysis of count data*. Cambridge University Press.
- [2] Chaney, S.E.P., Charity, M., David, M. and Aban, I. (2013). Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation Simulation* **83(9)**, 1671-1683.
- [3] Consul, P.C. and Famoye, F. (1992). Generalized Poisson regression model. *Communication in Statistics: Theory and Methods*, **21(1)**, 89–109.
- [4] Consul, P.C. and Jain, G.C. (1973). A Generalization of the Poisson distributions. *Technometrics* **15(4)**, 791-799.
- [5] Consul, P.C. (1989). *Generalized Poisson distributions: Properties and applications*, New York, Marcel Dekker.
- [6] Czado, C., Erhardt, V., Min, A., Wagner S. (2007). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. *Statistical Modelling* **7(2)**, 125-153.
- [7] Czado, C., Min, A., (2005). Consistency and asymptotic normality of the Maximum likelihood estimator in a zero-inflated generalized Poisson regression. Discussion paper, Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, No. 423.
- [8] Dean, C.B. (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of American Statistical Association* **87(418)**, 451–457.
- [9] Dean, C. and Lawless, J.F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of American Statistical Association* **84(406)**, 467–472.
- [10] Erhardt, V. (2011). ZIGP: zero-inflated generalized Poisson (ZIGP) regression models. R package, version 3.8.
- [11] Famoye, F. and Singh, K.P. (2006). Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. *Journal of Data Science* **4(1)**, 117-130.

- 
- [12] Famoye, F., Wulu, J. and Singh, K.P. (2004). On the Generalized Poisson Regression Model with an Application to Accident data. *Journal of Data Science* **2(3)**, 287-295.
- [13] Famoye, F. (1997). Parameter estimation for generalized negative binomial distribution. *Communication in Statistics: Simulation and Computation* **26(1)**, 269-279.
- [14] Famoye, F. (1993). Restricted generalized Poisson regression model. *Communication in Statistics: Theory and Methods* **22(5)**, 1335-1354.
- [15] Feller, W. (1943). On a general class of contagious distributions. *Annals of Mathematical Statistics* **14(4)**, 389-400.
- [16] Frome, E.L., Kutner, M.H. and Beauchamp, J.J. (1973). Regression analysis of Poisson distributed data. *Journal of American Statistical Association* **68(344)**, 935-940.
- [17] Gupta, P.L., Gupta, R.L. and Tripathi, R.C. (2004). Score test for Zero-Inflated Generalized Poisson Regression Model. *Communication in Statistics: Theory and Methods* **33(1)**, 47-64.
- [18] Gupta, P.L., Gupta, R.C. and Tripathi, R.C. (1996). Analysis of zero-adjusted count Data. *Computational Statistics and Data Analysis* **23(2)**, 207-218.
- [19] Gurmu, S., Rilstone, P. and Stern, S. (1999). Semiparametric estimation of count regression models. *Journal of Econometrics* **88(1)**, 123-150.
- [20] Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of applied econometrics* **12(3)**, 225-242.
- [21] Heibron, D.C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal* **36(5)**, 531-547.
- [22] Joe, H. and Zhu, R. (2005). Generalized Poisson distribution: The property of a mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal* **47(2)**, 219-229.
- [23] Lambert, D. (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* **34(1)**, 1-14.
- [24] Lawal, B.H. (2012). Zero-inflated count regression models with applications to some examples. *Quality quantity* **46(1)**, 19-38.
- [25] McCullough, P. and Nelder, J.A. (1989). *Generalized linear models*. Chapman and Hall.
- [26] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33(3)**, 341-365.
- [27] Neyman, J. (1939). On a new class of contagious distributions applicable in entomology and bacteriology. *Annals of Mathematical Statistics* **10(1)**, 35-57.
- [28] Puig, P. and Valero, J. (2006). Count Data Distributions: Some Characterizations with Applications. *Journal of American Statistical Association* **101(473)**, 332-340.

- 
- [29] Ridout, M., Hinde, J. and Demetrio, C.G.B. (2001). A score test for a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57(1)**, 219-223.
- [30] Santolino, M. and Boucher, J.P. (2009). Modeling the disability severity score in motor insurance claims: an application to the Spanish case. *Research Institute of Applied Economics Working Paper 2*, 25 pages.
- [31] Wagh Y.S. and Kamalja K.K. (2017a). Modelling Auto Insurance Claims in Singapore. *Sri Lankan Journal of Applied Statistics* (accepted for publication).
- [32] Wagh Y.S. and Kamalja K.K. (2017b). Comparison of methods of estimation for parameters of Generalized Poisson distribution through Simulation study. *Communications in Statistics-Simulation and Computation* **46(5)**, 4098-4112.
- [33] Wagh Y.S. and Kamalja K.K. (2017c). Zero-Inflated models and estimation in Zero-Inflated Poisson distribution, *Communications in Statistics-Simulation and Computation* (DOI-10.1080/03610918.2017.1341526).
- [34] Wang, W. and Famoye, F. (1997). Modelling Household Fertility Decision with Generalized Poisson Regression. *Journal of Population Economics* **10(3)**, 273-283.
- [35] Wulu, J.T, Singh, K.P., Famoye, F. and McGwin, G. (2002). Regression analysis of count data. *Journal of the Indian Society of Agricultural Statistics* **55(2)**, 220-231.
- [36] Xiang, L., Lee, A.H., Yau, K.K.W. and McLachlan, G.J. (2007). A score test for overdispersion in zero-inflated Poisson mixed regression model. *Statistics in medicine* **26(7)**, 1608–1622.
- [37] Yang, Z., Hardin, J.W. and Addy, C.L. (2010). Score test for zero-inflation in overdispersed count data. *Communication in Statistics: Theory and Methods* **39(11)**, 2008-2030.
- [38] Yang, Z., Hardin, J.W. and Addy, C.L. (2009). Testing overdispersion in the zero-inflated Poisson model. *Journal of Statistical Planning and Inference* **139(9)**, 3340-3353.
- [39] Yip, K.C.H. and Yau, K.K.W. (2005). On Modelling Claim Frequency Data in General Insurance with extra Zeros. *Insurance: Mathematics and Economics* **36(2)**, 153-163.
- [40] Zamani, H. and Ismail, N. (2012). Functional form for the generalized Poisson regression model. *Communication in Statistics: Theory and Methods* **41(20)**, 3666-3675.
- [41] Zamani, H. and Ismail, N. (2014). Functional form for the zero-inflated generalized Poisson regression model. *Communication in Statistics: Theory and Methods* **43(3)**, 515-529.
- [42] Zamani, H. and Ismail, N. (2013). Score test for testing zero-inflated Poisson regression against zero-inflated generalized Poisson alternatives. *Journal of Applied Statistics* **40(9)**, 2056-2068.