Pak.j.stat.oper.res. Vol.20 No. 3 2024 pp 369-381 DOI: http://dx.doi.org/10.18187/pjsor.v20i3.4538

Pakistan Journal of Statistics and Operation Research

# Robust parameter estimation for one-inflated positive Poisson Lindley distribution under the presence and absence of outliers with applications to crime data

Razik Ridzuan Mohd Tajuddin[1]*, Muhammad Aslam Mohd Safari[2], Noriszura Ismail[3]

* Corresponding Author

1. Department of Mathematical Sciences, Universiti Kebangsaan Malaysia, Malaysia, rrmt@ukm.edu.my
2. Department of Mathematics and Statistics, Universiti Putra Malaysia, Malaysia, aslam.safari@upm.edu.my
3. Department of Mathematical Sciences, Universiti Kebangsaan Malaysia, Malaysia, ni@ukm.edu.my

## Abstract

The one-inflated positive Poisson Lindley model has been recently introduced as an alternative in modelling positive count data with a large number of ones: a phenomenon known as one-inflation. In the presence of one-inflation, this model has a high tendency to be influenced by outliers, making usual parameter estimations to be less robust. Hence, several estimators: maximum likelihood, method of moments, ordinary least squares, weighted least squares, Cramér-Von Mises, modified Cramér-Von Mises (MCVM) and maximum product of spacing (MPS); for the parameters of the model are also proposed and investigated in terms of unbiasedness, consistency and joint efficiency under the presence and absence of outliers. When the outliers are absent, the MPS estimator is the best estimator and when the outliers are present, the MCVM estimator is the best estimator. Model fittings to two real datasets with one-inflation and outliers support the simulation results and conclude that the MCVM estimator is the best estimator. Based on the best robust estimator, the population size of the number of offenders as well as the likelihood of arrests were estimated.

**Key Words:** Excess Ones; Outliers; Population Size Estimator; Robust Estimator; Zero-Truncated Poisson Lindley.

**Mathematical Subject Classification:** 60E05; 62-08; 62F10; 62P25

## 1. Introduction

Truncating discrete count data distributions is one way of developing new distributions for modelling positive count data, resulting in zero-truncated distributions. However, recent studies have shown the importance of including a one-inflation parameter to cater for excess ones in the positive count data, yielding one-inflated zero-truncated distributions, such as one-inflated positive Poisson (Godwin & Böhning, 2017), one-inflated zero-truncated negative binomial (Godwin, 2017), one-inflated positive Poisson mixtures (Godwin, 2019) and one-inflated positive Poisson Lindley (Tajuddin et al., 2022a) distributions. The one-inflation parameter in the models captures the extra proportion of ones that are unexplained by the non-truncated discrete models. The large number of ones portrayed by the one-inflation parameter also contributes to dispersion in the data (Tajuddin et al., 2022a).

The effects of outliers were studied on zero-inflated null count data models (Yang et al., 2011) and zero-inflated regression count data models (Hall & Shen, 2010; Tüzen et al., 2020; Zandkarimi et al., 2019). The presence of outliers in the zero-inflated Poisson model is very influential since there is an abundance of zeros (Yang et al., 2011). For the zero-inflated regression model which is a special case of finite mixture models, the usual maximum likelihood estimator is highly sensitive when the outliers are present as well as when the mixture components are poorly separated (Hall & Shen, 2010; Zandkarimi et al., 2019). A comprehensive simulation study was conducted by Tüzen et al. (2020) on the effects of outliers and excess zeros in count data models, including non-inflated, zero-inflated and hurdle models. The authors found that the zero-inflated negative binomial and negative binomial hurdle models provide better fittings than other contending distributions when the effects of outliers and zero-inflation are varied.

To combat the issue of outliers in the count data, several robust estimators have been proposed and utilized in the literature. Since the median is not influenced by outliers, Coeurjolly and Rousseau Trepanier (2020) have developed a simple robust estimator for the median of a jittered Poisson distribution. The estimator is found to be consistent, asymptotically normal, and efficient for up to $10^9$ data samples. Other techniques in obtaining robust estimators found in the literature are trimmed mean, Winsorized mean, Tukey’s M-estimator and robust expectation-solution approaches. Yang et al. (2011) have employed both trimmed mean and Winsorized mean to obtain robust estimates of the parameters in the zero-inflated Poisson model. Although the estimators based on the trimmed mean are robust, removing outliers may lead to deformity in the estimators, and thus, Winsorized mean is a better alternative way of dealing with outliers (Yang et al., 2011). The resulting estimators based on trimmed mean and Winsorized mean perform better than the traditional maximum likelihood estimator under the presence of outliers. In conjunction with that, Li et al. (2011) have introduced a trimmed mean-kind estimator known as middle mean which only considers the middle of non-zero observations. The middle mean is equated with the non-zero part of the mean function.

Robust expectation-solution (RES) approach, which is a modification to the expectation-maximization (EM) algorithm, is another approach to deal with outliers. The RES approach adds robustified estimating functions in the M-step of the EM algorithm (Hall & Shen, 2010). A similar RES approach in estimating robust parameters has also been utilized to obtain robust estimates for parameters in a multilevel zero-inflated negative binomial model (Zandkarimi et al., 2019). Similar to the RES approach, another modified EM algorithm, known as the expectation-solution algorithm, was developed to investigate robust inference in the joint modeling of multilevel zero-inflated Poisson and Cox models with applications to health sciences (Zandkarimi et al., 2021).

The M-estimator and its variations using different objective functions have also been widely used as a robust estimator for count data models. For example, the properties of the robust M-estimator for Poisson and negative binomial models have been investigated (Cadigan & Chen, 2001). The M-estimator for the negative binomial model is better when the Poisson assumptions are uncertain (Cadigan & Chen, 2001). Recently, Tukey’s M-estimator for the Poisson parameter is modified so that data with small means could be investigated (Elsaied & Fried, 2016).

For one-inflated positive count data distributions, previous studies concerned with the performance of the population size estimator when the underlying distributions are one-inflated distributions (Godwin, 2017, 2019; Godwin & Böhning, 2017; Tajuddin et al., 2021, 2022a, 2022b). Tajuddin et al. (2021) investigated the performance of several estimators for the parameters of one-inflated positive Poisson distribution but not in the presence of outliers. Therefore, following the works of Yang et al. (2011) and Tüzen et al. (2020), this paper ultimately investigates the robustness of the estimators in the presence and absence of outliers for the one-inflated positive Poisson Lindley distribution, proposed by Tajuddin et al. (2022a).

For starters, following the work of Tajuddin et al. (2022a) on the one-inflated positive Poisson Lindley model, the estimators for the parameters of the one-inflated positive Poisson Lindley model based on maximum likelihood, method of moments, least squares, maximum product of spacing, Cramér-von Mises and modified Cramér-von Mises approaches will be discussed in this study. The estimators other than the method of moments involve optimization framework. The maximum likelihood estimator maximizes the likelihood function of the one-inflated positive Poisson Lindley model. The ordinary and the weighted least squares estimators minimize the least squares functions. The Cramér-von Mises and the modified Cramér-von Mises approaches minimize the distance function between theoretical and empirical distribution functions. Finally, the maximum product of spacing estimator minimizes the geometric mean of distance between two consecutive distribution functions. More details on these estimators can be found in Section 3.

Since the model has an abundance of ones, it is reasonable to believe that the conventional maximum likelihood and the moment estimators will be less robust in the presence of outliers. However, the remaining estimators are expected to be robust in the presence of outliers up to a certain degree (certain percentage of outliers). This hypothesis surely needs to be tested, and the results can be used in determining the best estimator in the presence of outliers. Hence, the unbiasedness and the consistency properties of all stated estimators will be investigated in the presence and absence of outliers via comprehensive simulation studies.

## 2. One-inflated positive Poisson Lindley (OIPPL) distribution and its statistical properties

The probability mass function for a random variable $Y$ which follows $OIPPL$ distribution with parameters $\omega$ and $\theta$ is given as (Tajuddin et al., 2022a):

$$\Pr\left(Y=y\mid\omega,\theta\right)=\begin{cases}\omega+\left(1-\omega\right)\dfrac{\theta^{2}\left(\theta+3\right)}{\left(\theta+1\right)\left(\theta^{2}+3\theta+1\right)}; y=1\\ \left(1-\omega\right)\dfrac{\theta^{2}\left(\theta+y+2\right)}{\left(\theta+1\right)^{y}\left(\theta^{2}+3\theta+1\right)} \quad ; y\geq 2,\end{cases} \tag{1}$$

where $\theta > 0$ and $0 < \omega < 1$ refers to the one-inflation parameter for $OIPPL$ distribution. The $OIPPL$ distribution can be described as a zero-truncated Poisson Lindley distribution with an additional inflating parameter for excess ones. The first two moments about origin, variance, and dispersion formulae of $OIPPL$ model are respectively given as:

$$E\left(Y\right)=\omega+\left(1-\omega\right)\frac{\left(\theta+1\right)^{2}\left(\theta+2\right)}{\theta\left(\theta^{2}+3\theta+1\right)}, \tag{2}$$

$$E\left(Y^{2}\right)=\omega+\left(1-\omega\right)\frac{\left(\theta+1\right)^{2}\left(\theta^{2}+4\theta+6\right)}{\theta^{2}\left(\theta^{2}+3\theta+1\right)}, \tag{3}$$

$$Var\left(Y\right)=\left(1-\omega\right)\left\{\frac{\left(\theta+1\right)^{2}\left(\theta^{3}+6\theta^{2}+10\theta+2\right)}{\theta^{2}\left(\theta^{2}+3\theta+1\right)^{2}}+\omega\left[\frac{\left(\theta+1\right)^{2}\left(\theta+2\right)}{\theta\left(\theta^{2}+3\theta+1\right)}-1\right]^{2}\right\}, \tag{4}$$

and

$$d=\left[\frac{\omega\rho_{z}}{1-\omega}+\frac{1}{d_{z}}\right]^{-1}+\left[E_{z}\left(Y\right)-1\right]^{2}\left[\frac{1}{1-\omega}+\frac{E_{z}\left(Y\right)}{\omega}\right]^{-1}, \tag{5}$$

where $\rho_Z = 1/Var_Z(Y)$, $d_Z = Var_Z(Y)/E_Z(Y)$, $E_Z(Y)$ and $Var_Z(Y)$ are the moment about the origin and the variance for the zero-truncated Poisson Lindley distribution, respectively. One can easily show that $Var(Y)$ takes the form $Var(Y) = (1-\omega)\{Var_Z(Y) + \omega[E_Z(Y) - 1]^2\}$.

## 3. Parameters estimation for OIPPL distribution

Parameter estimation for any distribution is a crucial step in modeling data. Therefore, several ways of estimating the parameters for $OIPPL$ distribution, which includes method of moments (MoM), maximum likelihood (ML), ordinary least squares (OLS), weighted least squares (WLS), maximum product spacing (MPS), Cramér-Von Mises (CVM) and modified Cramér-Von Mises (MCVM).

### *3.1 Method of Moments (MoM)*

By equating the first two sample moments with equations (2) and (3), two new equations are obtained and given as:

$$m_{1}=\tilde{\omega}+\left(1-\tilde{\omega}\right)\frac{\left(\tilde{\theta}+1\right)^{2}\left(\tilde{\theta}+2\right)}{\tilde{\theta}\left(\tilde{\theta}^{2}+3\tilde{\theta}+1\right)}, \tag{6}$$

and

$$m_{2}=\tilde{\omega}+\left(1-\tilde{\omega}\right)\frac{\left(\tilde{\theta}+1\right)^{2}\left(\tilde{\theta}^{2}+4\tilde{\theta}+6\right)}{\tilde{\theta}^{2}\left(\tilde{\theta}^{2}+3\tilde{\theta}+1\right)}, \tag{7}$$

where $m_j = \sum_{i=1}^{n} {y_i}^j / n = \sum_{y=1}^{\infty} \left(n_y y^j\right)/n$, $\tilde{\omega}$ and $\tilde{\theta}$ are the MoM estimators for $\omega$ and $\theta$, respectively. Solving the two equations above will yield a quintic equation:

$$A\tilde{\theta}^5 + B\tilde{\theta}^4 + C\tilde{\theta}^3 + D\tilde{\theta}^2 + E\tilde{\theta} + F = 0, \tag{8}$$

where $A = m_2 - 3m_1 + 2$, $B = 7m_2 - 23m_1 + 16$, $C = 15m_2 - 61m_1 + 46$, $D = 10m_2 - 68m_1 + 58$, $E = 2m_2 - 34m_1 + 32$ dan $F = 6(1 - m_1)$. Equation (8) can be solved numerically. The resulting $\tilde{\theta}$ is then substituted into equation (6) to obtain $\tilde{\omega}$.

### *3.2 Maximum Likelihood (ML)*

The log-likelihood function, $l$ for a random variable $Y$ which follows the $OIPPL$ model is given as:

$$l = \ln L(\omega, \theta) = \sum_{i=1}^{n} \ln f_O\left(y_i \mid \omega, \theta\right) = \sum_{y=1}^{k} n_y \ln f_O\left(y \mid \omega, \theta\right),$$

where $n_y$ refers to the number of $y$-valued observations. The ML estimators for $\omega$ and $\theta$ can be obtained by differentiating $l$ with respect to $\omega$ and $\theta$, given as (Tajuddin et al., 2022a):

$$\hat{\omega} = 1 - \frac{\left(n - n_1\right)}{n} \frac{\left(\hat{\theta} + 1\right)\left(\hat{\theta}^2 + 3\hat{\theta} + 1\right)}{\left(\hat{\theta}^2 + 4\hat{\theta} + 1\right)}, \tag{9}$$

and

$$\left(n - n_1\right)\left[\frac{\hat{\theta}\left(\hat{\theta} + 2\right)\left(\hat{\theta}^2 + 6\hat{\theta} + 3\right)}{\left(\hat{\theta}^2 + 4\hat{\theta} + 1\right)^2} + \frac{3\hat{\theta} + 2}{\hat{\theta}\left(\hat{\theta}^2 + 3\hat{\theta} + 1\right)}\right] + nm_1 - n_1 + \sum_{y=1}^{\infty} \frac{n_y}{\hat{\theta} + y + 2} = 0, \tag{10}$$

where $\hat{\omega}$ and $\hat{\theta}$ are the ML estimators for $\omega$ and $\theta$, respectively. Equation (10) can be solved numerically. Böhning and Ogden (2021) have provided a general estimator for the 'flation' parameter and with closer inspection, the $\hat{\omega}$ above falls into the general estimator with reparameterization of $w = 1 - \omega$ (Tajuddin et al., 2022a).

### *3.3 Ordinary and Weighted Least Squares*

For count data, suppose $y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(n)}$ is the order statistics of the data, which follows the $OIPPL$ model. It is known that:

$$E\left[F\left(Y_{(i)}\right)\right] = \frac{i}{n+1} \text{ and } Var\left[F\left(Y_{(i)}\right)\right] = \frac{i\left(n - i + 1\right)}{\left(n+1\right)^2\left(n+2\right)}; \quad i = 1, 2, \ldots, n.$$

However, for count data, the above mean and variance are best written with respect to the frequency of the data, given as:

$$E\left[F\left(Y_{n_y}\right)\right] = \frac{1}{n+1}\sum_{j=1}^{y} n_j \text{ and } Var\left[F\left(Y_{n_y}\right)\right] = \frac{\sum_{j=1}^{y} n_j \left(n - \sum_{j=1}^{y} n_j + 1\right)}{\left(n+1\right)^2\left(n+2\right)}; \quad y = 1, 2, 3, \ldots$$

The OLS estimates of parameters $\omega$ and $\theta$ can be obtained by minimizing the function:

$$O(\omega, \theta) = \sum_{y=1}^{k}\left[F(y) - \sum_{j=1}^{y} \frac{n_j}{n+1}\right]^2. \tag{11}$$

The WLS estimates of parameters $\omega$ and $\theta$ can be obtained by minimizing the function:

$$W(\omega, \theta) = \sum_{y=1}^{k}\left\{\frac{\left(n+1\right)^2\left(n+2\right)}{\sum_{j=1}^{y} n_j \left(n - \sum_{j=1}^{y} n_j + 1\right)}\left[F(y) - \sum_{j=1}^{y} \frac{n_j}{n+1}\right]^2\right\}. \tag{12}$$

### *3.4 Cramér-von Mises and modified Cramér-von Mises*

The estimators based on Cramér-von Mises (CVM) and modified Cramér-von Mises (MCVM) techniques involve minimizing the distance between the cumulative distribution function and the empirical distribution function. The estimators of parameters $\omega$ and $\theta$ using CVM and MCVM techniques can be obtained by minimizing the functions below, respectively, using R software (R Core Team, 2024).

$$C\left(\omega,\theta\right)=\frac{1}{12n}+\sum_{y=1}^{k}\left\{F\left(y\right)-\frac{1}{2n}\left[2\left(\sum_{j=1}^{y}n_{j}\right)-1\right]\right\}^{2}. \tag{13}$$

$$MC\left(\omega,\theta\right)=\sum_{y=1}^{k}\left\{\ln\left[F\left(y\right)^{-1}-1\right]-\ln\left[\frac{\left(n-1/2\right)}{\left(\sum_{j=1}^{y}n_{j}-3/4\right)}-1\right]\right\}^{2}. \tag{14}$$

*3.5 Maximum Product of Spacings*

The MPS estimator was first proposed by Cheng and Amin (1983) which considers the differences between consecutive values of cumulative distribution function. The uniform spacing for a random sample from the $OIPPL$ model can be defined as $D_y = F(y) - F(y-1)$, where $F(0) = 0$, $F(k+1) = 1$ and $\sum_{y=1}^{k} D_y = 1$. Both parameters $\omega$ and $\theta$ can be estimated by maximizing the geometric mean of $D_y$, given as:

$$G=\left[\prod_{y=1}^{k+1}D_{y}\right]^{\frac{1}{k+1}},$$

with respect to $\omega$ and $\theta$. Similarly, one can solve the log-geometric mean of $D_y$, given as:

$$g=\ln G=\frac{1}{k+1}\sum_{y=1}^{k+1}\ln D_{y}.$$

## 4. Comparison study of the estimators

The performance of the ML, MoM, OLS, WLS, CVM, MCVM and MPS estimators for the parameters of the $OIPPL$ are investigated in the aspect of unbiasedness and consistencies via simulation studies. The simulated data comes from the $OIPPL$ model with varying $\omega$ and $\theta$ values based on different values of population size, $N$. The pseudo-algorithms for generating simulated data as well as investigating the performance of each estimator are given as follows:

- Step 1 : Generate $N = 200$ data that follows Poisson Lindley distribution with $\theta = 1.0$.
- Step 2 : Remove $n_0$ zero-valued data to obtain $n$ positive count data.
- Step 3 : Replace $k$ data from $n$ positive count data into '1' counts with proportion $\omega = 0.3$ to exhibit the one-inflation property such that $k = \lfloor \omega n \rfloor$.
- Step 4 : Alter $p$ proportion of data with large numbers to mimic the presence of outliers, where $p = 0\%, 2\%, 4\%$.
- Step 5 : Fit the new data with $n$ data to $OIPPL$ model and estimate the values of $\omega$ and $\theta$ using ML, MoM, OLS, WLS, CVM, MCVM and MPS estimation techniques.
- Step 5 : Repeat steps 1-4 for 1000 times and get 1000 estimates for each parameter for all seven estimation techniques as well as compare these estimates with the true values of the parameters using mean absolute bias ($MAB$) and mean squared error ($MSE$), given respectively as:

  $$MAB=\frac{1}{1000}\sum_{i=1}^{1000}\left|\hat{\gamma}_{i}-\gamma\right|,$$

  and

  $$MSE=\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\gamma}_{i}-\gamma\right)^{2},$$

  where $\hat{\gamma}_i$ is the $i^{\text{th}}$ estimated value of $\gamma$ depending on the estimation technique used and $\gamma = (\theta, \omega)$.
- Step 6 : Repeat steps 1-5 using $\omega = 0.6$ as well as population size $N = 1000$.

The best estimator must result in the smallest $MAB$ and $MSE$. Since there are two estimated parameter values, the deficiency criterion, $Def = MSE_{\hat{\theta}} + MSE_{\hat{\omega}}$ (Akgül et al., 2016; Tajuddin et al., 2021) will be used. The best estimator must also yield the smallest $Def$ value. The simulation results are tabulated in Table 1 – Table 4.

Table 1 shows the $MAB$ and the $MSE$ values for various estimation techniques when a total of $N = 200$ data generated from the OIPPL distribution with $\theta = 1.0$ and $\omega = 0.3, 0.6.$ under the presence of outliers ($p = 2\%, 4\%$)

and in the absence of outliers ($p = 0\%$), whereas Table 2 shows the associated $Def$ values. From Table 1 and Table 2, when the outliers are absent ($p = 0\%$), the MPS estimator showed the smallest $MAB$, $MSE$ and $Def$ values. On the other hand, when the outliers are present ($p = 2\%, 4\%$), the MCVM estimator provided the smallest $MAB$, $MSE$ and $Def$ values, suggesting that the MCVM estimator is the most desirable estimator. Furthermore, the MoM estimator provided the largest values of $MAB$, $MSE$ and $Def$, making it the least desirable estimator.

Table 1 The $MAB$ and the $MSE$ values of the estimators under various estimation techniques when $N = 200, \theta = 1.0$ and several values of $\omega$ and $p$.

| $\omega$ | $p$ | Estimators | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ML | | MOM | | OLS | | WLS | | CVM | | MCVM | | MPS | |
| | | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ |
| 0.3 | 0% | 0.1328<br>0.0310 | 0.0693<br>0.0080 | 0.1652<br>0.0470 | 0.1027<br>0.0169 | 0.1299<br>0.0271 | 0.0678<br>0.0073 | 0.1374<br>0.0294 | 0.0774<br>0.0094 | 0.1308<br>0.0287 | 0.0691<br>0.0078 | 0.1384<br>0.0330 | 0.0703<br>0.0082 | **0.1296**<br>**0.0258** | **0.0649**<br>**0.0067** |
| | 2% | 0.1699<br>0.0370 | 0.0649<br>0.0063 | 0.2805<br>0.0863 | 0.1423<br>0.0257 | 0.1687<br>0.0384 | 0.0681<br>0.0071 | 0.2194<br>0.0606 | 0.0936<br>0.0127 | 0.1540<br>0.0329 | 0.0664<br>0.0068 | **0.1428**<br>**0.0298** | **0.0665**<br>**0.0068** | 0.2273<br>0.0600 | 0.0700<br>0.0073 |
| | 4% | 0.3048<br>0.0982 | 0.0805<br>0.0092 | 0.4368<br>0.1938 | 0.2025<br>0.0456 | 0.2684<br>0.0830 | 0.0843<br>0.0103 | 0.3591<br>0.1395 | 0.1401<br>0.0249 | 0.2480<br>0.0724 | 0.0790<br>0.0092 | **0.1982**<br>**0.0500** | **0.0663**<br>**0.0066** | 0.3452<br>0.1244 | 0.0861<br>0.0102 |
| 0.6 | 0% | 0.1805<br>0.0590 | 0.0508<br>0.0045 | 0.2182<br>0.0895 | 0.0757<br>0.0096 | 0.1783<br>0.0494 | 0.0488<br>0.0040 | 0.1849<br>0.0539 | 0.0555<br>0.0051 | 0.1787<br>0.0536 | 0.0503<br>0.0044 | 0.1855<br>0.0604 | 0.0510<br>0.0047 | **0.1753**<br>**0.0459** | **0.0465**<br>**0.0036** |
| | 2% | 0.2411<br>0.0693 | 0.0474<br>0.0034 | 0.3599<br>0.1353 | 0.0916<br>0.0107 | 0.2422<br>0.0746 | 0.0506<br>0.0039 | 0.3084<br>0.1138 | 0.0689<br>0.0068 | 0.2177<br>0.0624 | 0.0500<br>0.0039 | **0.1986**<br>**0.0542** | **0.0499**<br>**0.0039** | 0.3099<br>0.1069 | 0.0485<br>0.0035 |
| | 4% | 0.4031<br>0.1676 | 0.0483<br>0.0035 | 0.5030<br>0.2551 | 0.1025<br>0.0124 | 0.3707<br>0.1535 | 0.0540<br>0.0043 | 0.4583<br>0.2230 | 0.0847<br>0.0094 | 0.3431<br>0.1345 | 0.0520<br>0.0040 | **0.2965**<br>**0.1049** | **0.0472**<br>**0.0034** | 0.4295<br>0.1912 | 0.0475<br>0.0033 |

The best estimator is written in bold.

Table 2 The $Def$ values of the estimators under various estimation techniques $N = 200, \theta = 1.0$ and several values of $\omega$ and $p$.

| $\omega$ | $p$ | Deficiency values | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | ML | MOM | OLS | WLS | CVM | MCVM | MPS |
| 0.3 | 0% | 0.0390 | 0.0639 | 0.0344 | 0.0388 | 0.0365 | 0.0412 | **0.0325** |
| | 2% | 0.0433 | 0.1120 | 0.0455 | 0.0733 | 0.0397 | **0.0366** | 0.0673 |
| | 4% | 0.1074 | 0.2394 | 0.0933 | 0.1644 | 0.0816 | **0.0566** | 0.1346 |
| 0.6 | 0% | 0.0635 | 0.0991 | 0.0534 | 0.0590 | 0.0580 | 0.0651 | **0.0495** |
| | 2% | 0.0727 | 0.1460 | 0.0785 | 0.1206 | 0.0663 | **0.0581** | 0.1104 |
| | 4% | 0.1711 | 0.2675 | 0.1578 | 0.2324 | 0.1385 | **0.1083** | 0.1945 |

The best estimator is written in bold.

The results of the simulation study for $N = 1000$ are summarized in Table 3 and Table 4. From Table 3 and Table 4, when the outliers are absent ($p = 0\%$), both ML and MPS estimators provided the smallest $MAB$, $MSE$ and $Def$ values for $\omega = 0.3$. However, when $\omega = 0.6$, the MPS estimator provided the smallest $MAB$, $MSE$ and $Def$ values. It is worth noting that when $\omega = 0.3$, the MPS estimator provided the second smallest values of $MAB$, $MSE$ and $Def$. Similar observations to Table 1 and Table 2 can be made when the outliers are present, in which the MCVM estimator provided the smallest $MAB$, $MSE$ and $Def$ values, whereas the MoM estimator provided the largest $MAB$, $MSE$ and $Def$ values.

Table 3 The $MAB$ and the $MSE$ values of the estimators under various estimation techniques when $N = 1000, \theta = 1.0, \omega = 0.3, 0.6$ and $p = 0\%, 2\%, 4\%$.

| $\omega$ | $p$ | Estimators | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ML | | MOM | | OLS | | WLS | | CVM | | MCVM | | MPS | |
| | | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ | $\hat{\theta}$ | $\hat{\omega}$ |
| 0.3 | 0% | **0.0586**<br>**0.0054** | **0.0299**<br>**0.0014** | 0.0705<br>0.0078 | 0.0433<br>0.0030 | 0.0604<br>0.0057 | 0.0303<br>0.0014 | 0.0649<br>0.0064 | 0.0350<br>0.0019 | 0.0602<br>0.0058 | 0.0305<br>0.0014 | 0.0628<br>0.0064 | 0.0310<br>0.0015 | 0.0594<br>0.0054 | 0.0295<br>0.0014 |
| | 2% | 0.1970<br>0.0408 | 0.0512<br>0.0034 | 0.3262<br>0.1077 | 0.1663<br>0.0290 | 0.1554<br>0.0276 | 0.0483<br>0.0032 | 0.2629<br>0.0721 | 0.1168<br>0.0151 | 0.1496<br>0.0259 | 0.0464<br>0.0030 | **0.1131**<br>**0.0164** | **0.0327**<br>**0.0016** | 0.2136<br>0.0476 | 0.0552<br>0.0038 |
| | 4% | 0.3274<br>0.1081 | 0.0800<br>0.0072 | 0.4571<br>0.2094 | 0.2127<br>0.0461 | 0.2906<br>0.0868 | 0.0898<br>0.0091 | 0.4125<br>0.1716 | 0.1794<br>0.0333 | 0.2854<br>0.0838 | 0.0875<br>0.0087 | **0.2135**<br>**0.0488** | **0.0471**<br>**0.0031** | 0.3377<br>0.1150 | 0.0823<br>0.0075 |
| 0.6 | 0% | 0.0755<br>0.0092 | 0.0225<br>0.0008 | 0.0933<br>0.0142 | 0.0330<br>0.0018 | 0.0763<br>0.0090 | 0.0223<br>0.0008 | 0.0841<br>0.0107 | 0.0277<br>0.0012 | 0.0765<br>0.0092 | 0.0225<br>0.0008 | 0.0784<br>0.0098 | 0.0227<br>0.0008 | **0.0750**<br>**0.0086** | **0.0218**<br>**0.0007** |
| | 2% | 0.2841<br>0.0828 | 0.0359<br>0.0017 | 0.4062<br>0.1662 | 0.1036<br>0.0113 | 0.2507<br>0.0675 | 0.0394<br>0.0021 | 0.3656<br>0.1370 | 0.0875<br>0.0084 | 0.2418<br>0.0632 | 0.0380<br>0.0020 | **0.2044**<br>**0.0470** | **0.0285**<br>**0.0012** | 0.2992<br>0.0919 | 0.0370<br>0.0018 |
| | 4% | 0.4282<br>0.1842 | 0.0425<br>0.0022 | 0.5202<br>0.2710 | 0.1058<br>0.0116 | 0.4252<br>0.1833 | 0.0626<br>0.0044 | 0.4942<br>0.2457 | 0.0991<br>0.0104 | 0.4186<br>0.1778 | 0.0612<br>0.0043 | **0.3645**<br>**0.1360** | **0.0400**<br>**0.0021** | 0.4322<br>0.1880 | 0.0419<br>0.0021 |

The best estimator is written in bold.

Table 4 The $Def$ values of the estimators under various estimation techniques $N = 1000, \theta = 1.0$ and several values of $\omega$ and $p$.

| $\omega$ | $p$ | Deficiency values | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | ML | MOM | OLS | WLS | CVM | MCVM | MPS |
| 0.3 | 0% | **0.0068** | 0.0108 | 0.0071 | 0.0083 | 0.0072 | 0.0079 | **0.0068** |
| | 2% | 0.0442 | 0.1367 | 0.0308 | 0.0872 | 0.0289 | **0.0180** | 0.0514 |
| | 4% | 0.1153 | 0.2555 | 0.0959 | 0.2049 | 0.0925 | **0.0519** | 0.1225 |
| 0.6 | 0% | 0.0100 | 0.0160 | 0.0098 | 0.0119 | 0.0100 | 0.0106 | **0.0093** |
| | 2% | 0.0845 | 0.1775 | 0.0696 | 0.1454 | 0.0642 | **0.0482** | 0.0937 |
| | 4% | 0.1864 | 0.2826 | 0.1877 | 0.2561 | 0.1821 | **0.1381** | 0.1901 |

The best estimator is written in bold.

In conclusion, regardless of the values of $\omega$ and $N$, when the outliers are absent, the MPS estimator is the best estimator for parameters $\theta$ and $\omega$. However, when the outliers are present, the MCVM estimator is the best estimator for parameters $\theta$ and $\omega$. For both values of $N$ and when the outliers are present, the MoM estimator is the worst estimator. This is expected because the MoM estimator is derived from the sample moments and sample moments are heavily influenced by outliers.

## 5. Applications

To illustrate and investigate the performance of the proposed estimators in fitting real data, two datasets are considered in this study. The data are tested for a large number of excess ones based on the one-inflation index proposed by Tajuddin et al. (2021). The formula for the index, as given by Tajuddin et al. (2021):

$$oi_{pp} = 1 + \frac{\ln(p_1)}{\ln[\exp(d+\mu-1)-1] - \ln(d+\mu-1)},$$

where $p_1$ refers to the proportion of one-valued data, $d$ refers to the dispersion index and $\mu$ refers to the mean. When dealing with sample data, one may use sample proportion, sample dispersion and sample mean to calculate a sample one-inflation index (Tajuddin et al., 2021). A positive index value indicates the presence of large one-valued observations in the dataset (Tajuddin et al., 2021). Additionally, the score test proposed by Godwin and Böhning (2017) can be used to investigates the presence of one-inflation via hypothesis testing with comparison to positive Poisson distribution.

Several statistics such as the chi-square goodness-of-fit test, the root mean squared error (RMSE), the mean absolute error (MAE), the root mean squared logarithmic error (RMSLE) and the median absolute deviation (MedAD) are used to determine the best model. Let $e_x = n_x - \hat{n}_x$, where $\hat{n}_x$ is the estimated frequency for each $n_x$, be the error value for $x$-valued data, then the formulae for RMSE, MAE, RMSLE and MedAD are respectively given as:

$$RMSE = \sqrt{\frac{1}{h}\sum_{x=1}^{h} e_x^{\ 2}},$$

$$MAE = \frac{1}{h}\sum_{x=1}^{h} \left| e_x \right|,$$

$$RMSLE = \sqrt{\frac{1}{h}\sum_{x=1}^{h} \left[\ln\left(n_x + 1\right) - \ln\left(\hat{n}_x + 1\right)\right]^2},$$

$$MedAD = med\left[e_x - med\left(e_x\right)\right],$$

where $h$ refers to the number of data groups and $med(s)$ refers to the median of data with $s$ observations. Generally, estimators that give model fitting with adequate goodness-of-fit and the smallest RMSE, MAE, RMSLE and MedAD values, are selected as the best estimators. For data with outliers, both RMSLE and MedAD are robust.

### *5.1 Example I*

The first dataset refers to the counts of prostitution arrests in Vancouver (Rossmo & Routledge, 1990). The data are given as $n_1 = 541$, $n_2 = 169$, $n_3 = 95$, $n_4 = 37$, $n_5 = 21$ and $n_6 = 23$, with sample one-inflation index of 0.4486, indicating the presence of excess ones in the data. The histogram and the boxplot of the counts of prostitution arrests data are provided in Figure 1. From Figure 1, the histogram suggests that the prostitution arrests data is skewed to the right and the boxplot suggests that there are three outliers: $n_4$, $n_5$ and $n_6$. Since the data have large number of ones and several outliers, the OIPPL distribution with different estimation techniques can be used for model fitting (see Section 5.3).

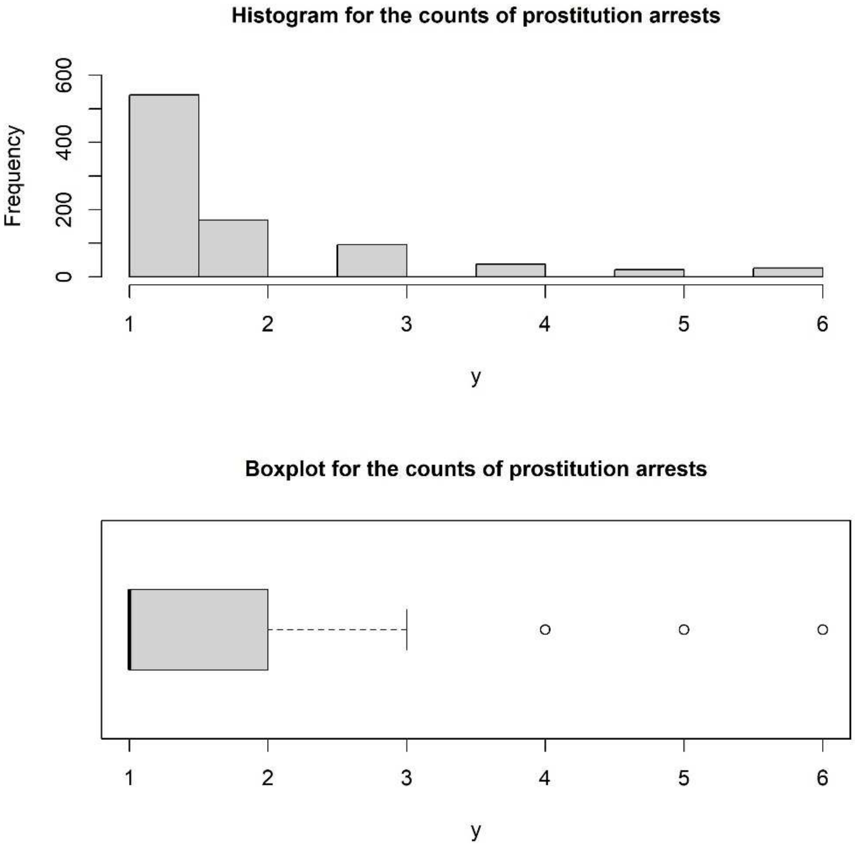

Figure 1 Histogram and the boxplot of the counts of prostitution arrests data

*5.2 Example II*

The second dataset refers to the counts of drunk drivers in the police records (Van Der Heijden et al., 2003). The data are given as $n_1 = 8877$, $n_2 = 481$, $n_3 = 52$, $n_4 = 8$ and $n_5 = 1$, with sample one-inflation index of 0.1543, suggesting the presence of excess ones in the data. The histogram and the boxplot of the counts of drunk drivers in the police records. From Figure 2, the histogram suggests that the drunk drivers data is skewed to the right and the boxplot suggests that there are four outliers: $n_2$, $n_3$, $n_4$ and $n_5$. Similar to Example I, the OIPPL distribution with different estimation techniques can be used for model fitting (see Section 5.3).

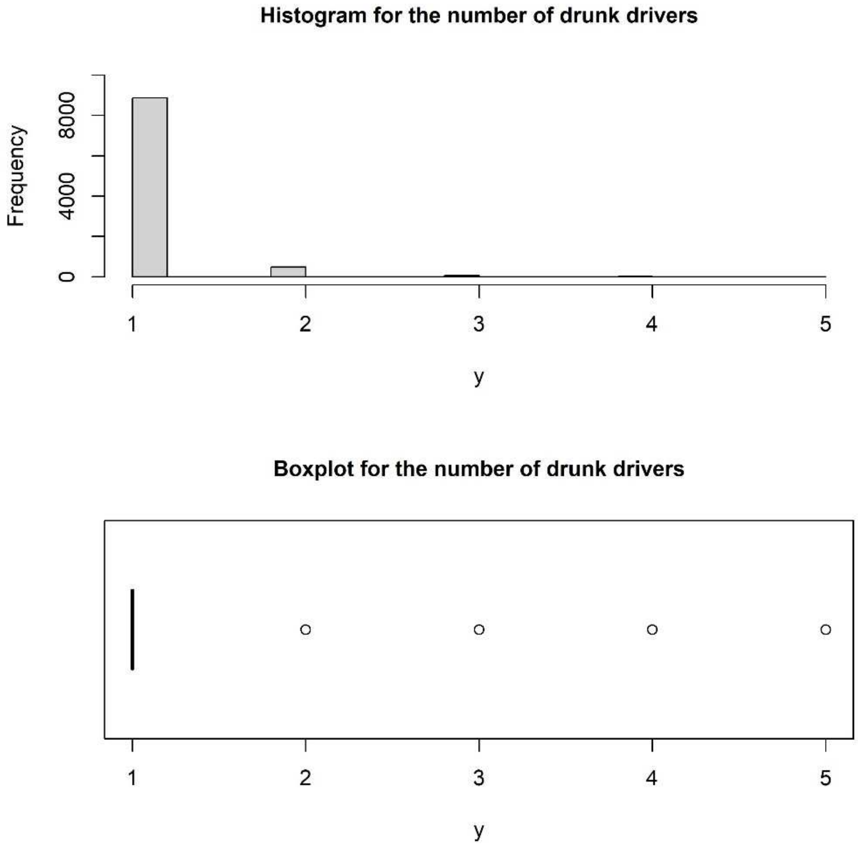

Figure 2 Histogram and the boxplot of the counts of drunk drivers in the police records

*5.3 Model Fittings*

All seven estimators: ML, MoM, OLS, WLS, CVM, MCVM and MPS; are used to estimate the two parameters of the OIPPL distribution based on the datasets described in Section 5.1 and Section 5.2. The summaries of the model fittings are given in Table 5.

For the first dataset in Table 5, every model fitting provides an adequate fit to the data except for WLS estimator. Therefore, the WLS estimator can be excluded in the list of the best models. Additionally, the model fitting resulting from MCVM estimator provides the smallest RMSE, MAE and MedAD values while the model fitting resulting from using OLS estimator provides the smallest RMSLE value. Moreover, its RMSE, MAE and MedAD values are very close to those from MCVM estimator. The RMSLE value from CVM estimator is comparable to those from MCVM and OLS estimators as well. Other estimators provide less desirable error values. Despite MCVM and OLS estimators producing close RMSE, MAE, RMSLE and MedAD values, the best estimation technique in fitting data with a large number of ones and outliers cannot be decided objectively.

For the second dataset in Table 5, all model fittings provide an adequate fit to the data. However, the model fitting resulting from the MCVM estimator provides the smallest RMSE and MAE values but also the largest RMSLE value. Conversely, the model fitting resulting from the WLS estimator provides the smallest RMSLE value but the largest RMSE and MAE values. The OLS estimator, on the other hand, provides the smallest MedAD value. Similar to the issue faced in Dataset I, an objective decision regarding the best estimation technique in the presence of one-inflation and outliers cannot be obtained.

Table 5 Summary of model fittings for prostitute arrests in Vancouver (Dataset I) and drunk drivers (Dataset II)

| Dataset | Estimation Techniques | | | | | | |
|---|---|---|---|---|---|---|---|
| | ML | MoM | OLS | WLS | CVM | MCVM | MPS |
| I | | | | | | | |
| $\hat{\theta}$ | 1.3653 | 1.4822 | 1.3120 | 1.6436 | 1.3183 | 1.3128 | 1.3476 |
| $\hat{\omega}$ | 0.2300 | 0.1552 | 0.2454 | 0.0917 | 0.2433 | 0.2463 | 0.2346 |
| RMSE | 12.163 | 28.787 | 10.751 | 40.541 | 10.857 | **10.698** | 11.577 |
| MAE | 4.074 | 9.274 | 3.594 | 12.262 | 3.668 | **3.481** | 3.893 |
| RMSLE | 0.114 | 0.171 | **0.095** | 0.272 | 0.096 | 0.096 | 0.106 |
| MedAD | 5.373 | 6.376 | 2.850 | 6.622 | 3.200 | **2.778** | 4.686 |
| $\chi^2$ | 2.875 | 7.493 | 2.189 | 17.406 | 2.232 | 2.193 | 2.566 |
| df | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| p-value | 0.4113 | 0.0577 | 0.5341 | 0.0006 | 0.5257 | 0.5333 | 0.4635 |
| II | | | | | | | |
| $\hat{\theta}$ | 8.3238 | 8.1526 | 8.3549 | 7.7998 | 8.4341 | 8.4759 | 7.9087 |
| $\hat{\omega}$ | 0.5067 | 0.5175 | 0.5040 | 0.5375 | 0.4999 | 0.4978 | 0.5291 |
| RMSE | 4.403 | 6.463 | 4.106 | 9.820 | 3.668 | **3.466** | 7.393 |
| MAE | 1.835 | 2.894 | 1.745 | 4.030 | 1.510 | **1.385** | 3.018 |
| RMSLE | 0.102 | 0.090 | 0.104 | **0.073** | 0.111 | 0.115 | 0.076 |
| MedAD | 1.012 | 2.855 | **0.537** | 4.432 | 0.707 | 0.881 | 1.699 |
| $\chi^2$ | 0.662 | 0.690 | 0.671 | 0.900 | 0.693 | 0.712 | 0.793 |
| df | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| p-value | 0.4159 | 0.4062 | 0.4127 | 0.3428 | 0.4051 | 0.3988 | 0.3732 |

The smallest value is written in bold.

Based on the two model fittings, the best estimation technique cannot be concluded because each criterion provides a different assessment. On that account, we employ an integrated approach which involves multiplicative aggregation through standard normal cumulative distribution function transformation (Masseran, 2018). The integrated approach combines all criteria involved in this study to produce an objective decision and subsequently selects the best model. Masseran (2018) integrated several statistics: Kolmogorov-Smirnov, Akaike Information Criterion, deviation of skewness and kurtosis and complementary of $R^2$; standardized them to scores and transformed them to a normal distribution, which takes values between 0 and 1. Each transformed score for the same distribution gets multiplied and the model that yields the smallest final multiplicative aggregation value is selected as the best model (Masseran, 2018).

This paper adopts the approach by Masseran (2018) by integrating the RMSE, MAE, RMSLE and MedAD values in the multiplicative aggregation. By doing so, the best estimation technique in fitting the datasets with one-inflation and outliers can be obtained. The transformed scores and their multiplicative aggregation values as well as the rank from the best to worst model fitting are provided in Table 6.

Table 6 Final transformed scores, multiplicative aggregation * values and their corresponding rank.

| Estimation techniques | Dataset I | | | | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | RMSLE | MedAD | Multiplicative aggregation * | Rank |
| ML | 0.005050 | 0.026990 | 0.138444 | 0.226891 | 0.004281 | 5 |
| MoM | 0.023545 | 0.227567 | 0.634824 | 0.368923 | 1.254863 | 6 |
| OLS | 0.004368 | 0.020878 | **0.058830** | 0.036600 | 0.000196 | 2 |
| WLS | 0.058147 | 0.472587 | 0.998025 | 0.407856 | 11.185536 | 7 |

| | | | | | | |
|---|---|---|---|---|---|---|
| CVM | 0.004416 | 0.021736 | 0.061835 | 0.049782 | 0.000295 | 3 |
| MCVM | **0.004344** | **0.019624** | 0.061835 | **0.034278** | **0.000181** | **1** |
| MPS | 0.004756 | 0.024529 | 0.098817 | 0.150806 | 0.001739 | 4 |

| Estimation techniques | Dataset II | | | | | |
|---|---|---|---|---|---|---|
| | RMSE | MAE | RMSLE | MedAD | Multiplicative aggregation * | Rank |
| ML | 0.004862 | 0.054260 | 0.123269 | 0.087097 | 0.002832 | 4 |
| MoM | 0.013246 | 0.178487 | 0.116709 | 0.304809 | 0.084103 | 6 |
| OLS | 0.004165 | 0.048158 | 0.124387 | **0.057344** | 0.001431 | 3 |
| WLS | 0.052449 | 0.425556 | **0.107846** | 0.585188 | 1.408616 | 7 |
| CVM | 0.003301 | 0.034784 | 0.128356 | 0.066908 | 0.000986 | 2 |
| MCVM | **0.002959** | 0.029019 | 0.130663 | 0.077929 | **0.000874** | **1** |
| MPS | 0.020015 | **0.003067** | 0.999328 | 0.072194 | 0.004428 | 5 |

* Multiplied by 1000 to scale up the small values.
The smallest value is written in bold.

Based on Table 6, the final multiplicative aggregation values of the MCVM estimator are the smallest for both datasets, indicating that the MCVM estimator is the best estimator when estimating parameters for the OIPPL distribution when the data have a large number of ones and outliers. This result is supported by the simulation results given in Table 1 – Table 4. On the other hand, the final multiplicative aggregation values of the WLS estimator are the largest for both datasets, which suggests that the WLS estimator is the worst estimator in estimating parameters for the OIPPL distribution when one-inflation and outliers are present. Therefore, if the data under consideration have a large number of ones and outliers are present, it is recommended to use the MCVM estimator over other estimators in estimating the parameters of the OIPPL distribution.

Based on the estimated parameters from using the MCVM technique, the unobserved number of uncaught prostitutes and drunk drivers can be estimated. Subsequently, the total number of offenders (both caught and uncaught) can be estimated. The total number of offenders for both prostitutes and drunk drivers can be estimated using (Tajuddin et al., 2022a):

$$\hat{N} = \frac{n\left(\hat{\theta}+1\right)^3}{\hat{\theta}^2 + 3\hat{\theta}+1},$$

with its estimated variance $\widehat{Var}(\widehat{N}) = A(\hat{\theta}) + B(\hat{\theta})C(\hat{\theta})$ (Tajuddin et al., 2022b), where

$$A\left(\hat{\theta}\right) = \frac{n\hat{\theta}^2\left(\hat{\theta}+2\right)\left(\hat{\theta}+1\right)^3}{\left(\hat{\theta}^2+3\hat{\theta}+1\right)^2},$$

$$B\left(\hat{\theta}\right) = \left[\frac{n\hat{\theta}\left(\hat{\theta}+4\right)\left(\hat{\theta}+1\right)^2}{\left(\hat{\theta}^2+3\hat{\theta}+1\right)^2}\right]^2,$$

$$C\left(\hat{\theta}\right) = \frac{1}{n}\left[-\frac{\hat{\theta}^3-2\hat{\theta}^2-4\hat{\theta}-2}{\hat{\theta}^2\left(\hat{\theta}+1\right)^3} + \frac{\hat{\theta}^2}{\left(\hat{\theta}+1\right)^2}\int_0^1 \frac{t^{\hat{\theta}+1}}{\hat{\theta}+1-t}\,dt\right]^{-1}.$$

Knowing the estimated values and the estimated variance, the 95% confidence interval can be obtained. Subsequently, the likelihood of arrests, $LOA$ can be obtained using (Chainey & Lazarus, 2021):

$$LOA = \frac{\text{total number of individual offenders arrested}}{\text{total number of estimated individuals}}.$$

Table 7 summarizes the estimated population size and its 95% confidence interval as well as the likelihood of arrests for both datasets. From Table 7, it can be concluded that less than half of the prostitutes have not been arrested yet whereas almost 90% of drunk drivers have not get caught by the police yet. These $LOA$ values will surely help the authority to be more alert and active in scouting the streets and conducting frequent police stops.

Table 7 Estimated population size with its corresponding 95% confidence interval (lower and upper values) and the likelihood of arrest.

| Dataset | $\hat{\theta}$ | $\hat{\omega}$ | $n$ | $\hat{N}$ | 95% confidence interval | | $LOA$ |
|---|---|---|---|---|---|---|---|
| | | | | | Lower | Upper | |
| I | 1.3128 | 0.2463 | 886 | 1645 | 1541 | 1749 | 53.86% |
| II | 8.4759 | 0.4978 | 9419 | 81, 555 | 80, 006 | 83, 104 | 11.55% |

## 6. Conclusions

This paper aimed to investigate the performance of several parameter estimation techniques for the one-inflated positive Poisson Lindley distribution in the presence and absence of outliers. From the simulation studies, it was found that the modified Cramér-Von Mises estimator provided the smallest error values in the aspect of unbiasedness and consistency when the data have outliers. In the absence of outliers, the maximum product spacing estimator yielded the smallest error values. For model fittings to real data with a large number of ones and outliers, since several criteria were used, an objective decision on the best estimator could not be obtained. With the help of an integrated approach that combines the information from the criteria: root mean squared error, mean absolute bias, root mean squared logarithmic error and median absolute deviation; the modified Cramér-Von Mises estimator was selected as the best estimator in estimating parameters for the one-inflated positive Poisson Lindley distribution when the data are one-inflated and have outliers. Based on the best estimator, the population size as well as the likelihood of arrests were calculated. It is believed that the likelihood of arrests for both prostitutes and drunk drivers datasets will inform the authorities to be more vigilant and active in capturing these offenders.

## Declarations

The authors are thankful for the financial assistance received from Universiti Kebangsaan Malaysia in the form of grant Geran Galakan Penyelidik Muda (grant no. GGPM-2023-052).

## References

Akgül, F. G., Şenoğlu, B., & Arslan, T. (2016). An alternative distribution to Weibull for modeling the wind speed data: Inverse Weibull distribution. *Energy Conversion and Management*, *114*, 234-240.

Böhning, D., & Ogden, H. E. (2021). General flation models for count data. *Metrika*, *84*(2), 245-261.

Chainey, S. P., & Lazarus, D. L. (2021). More Offenders, More Crime: Estimating the Size of the Offender Population in a Latin American Setting. *Social Sciences*, *10*(9), 348.

Cheng, R., & Amin, N. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*(3), 394-403.

Godwin, R. T., & Böhning, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *66*(2), 425-448.

Masseran, N. (2018). Integrated approach for the determination of an accurate wind-speed distribution model. *Energy Conversion and Management*, *173*, 56-64.

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. In R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rossmo, D. K., & Routledge, R. (1990). Estimating the size of criminal populations. *Journal of quantitative criminology*, *6*, 293-314.

Tajuddin, R. R. M., Ismail, N., & Ibrahim, K. (2021). Comparison of estimation methods for one-inflated positive Poisson distribution. *Journal of Taibah University for Science*, *15*(1), 869-881.

Tajuddin, R. R. M., Ismail, N., & Ibrahim, K. (2022a). Estimating Population Size of Criminals: A New Horvitz–Thompson Estimator under One-Inflated Positive Poisson–Lindley Model. *Crime & Delinquency*, *68*(6-7), 1004-1034.

Tajuddin, R. R. M., Ismail, N., & Ibrahim, K. (2022b). On Variance Estimation for the Population Size Estimator under One-Inflated Positive Poisson Distribution. *Malaysian Journal of Fundamental and Applied Sciences*, *18*(2), 237-244.

Van Der Heijden, P. G., Cruyff, M., & Van Houwelingen, H. C. (2003). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, *57*(3), 289-304.