

Robust Estimation for Zero-Inflated Poisson Regression

DANIEL B. HALL

Department of Statistics, University of Georgia

JING SHEN

Merial Limited

ABSTRACT. The zero-inflated Poisson regression model is a special case of finite mixture models that is useful for count data containing many zeros. Typically, maximum likelihood (ML) estimation is used for fitting such models. However, it is well known that the ML estimator is highly sensitive to the presence of outliers and can become unstable when mixture components are poorly separated. In this paper, we propose an alternative robust estimation approach, robust expectation-solution (RES) estimation. We compare the RES approach with an existing robust approach, minimum Hellinger distance (MHD) estimation. Simulation results indicate that both methods improve on ML when outliers are present and/or when the mixture components are poorly separated. However, the RES approach is more efficient in all the scenarios we considered. In addition, the RES method is shown to yield consistent and asymptotically normal estimators and, in contrast to MHD, can be applied quite generally.

Key words: expectation-maximization (EM) algorithm, excess zeros, expectation-solution algorithm, minimum Hellinger distance, outliers, robustness

1. Introduction

Count data with many zeros in addition to large non-zero values are common in a wide variety of disciplines. This phenomenon can be handled by a two-component mixture where one of the components is taken to be a degenerate distribution, having mass one at zero. The other component is a non-degenerate distribution such as the Poisson, binomial, negative binomial or other form depending on the situation. Lambert (1992) proposed the zero-inflated Poisson (ZIP) regression model and illustrated it through an analysis of data related to the incidence of manufacturing defects. In ZIP regression, the response vector is $\mathbf{y} = (y_1, \dots, y_n)^T$, where y_i is the observed value of the random variable Y_i . The Y_i s are assumed independent, where

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{Poisson}(\mu_i), & \text{with probability } 1 - p_i. \end{cases}$$

Moreover, the parameters $\mathbf{p} = (p_1, \dots, p_n)^T$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T$ are modelled through canonical link generalized linear models (GLM) as $\text{logit}(\mathbf{p}) = \mathbf{G}\boldsymbol{\gamma}$ and $\log(\boldsymbol{\mu}) = \mathbf{B}\boldsymbol{\beta}$, where $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are regression parameters, and \mathbf{G} and \mathbf{B} are corresponding design matrices that pertain to the probability of the zero state and the Poisson mean, respectively. The log-likelihood function for this model can be written as:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}) = & \sum_{y_i=0} \log\{e^{G_i^T \boldsymbol{\gamma}} + \exp(-e^{B_i^T \boldsymbol{\beta}})\} + \sum_{y_i>0} (y_i B_i^T \boldsymbol{\beta} - e^{B_i^T \boldsymbol{\beta}}) \\ & - \sum_{y_i>0} \log(y_i!) - \sum_{i=1}^n \log(1 + e^{G_i^T \boldsymbol{\gamma}}), \end{aligned} \quad (1)$$

where \mathbf{B}_i^\top and \mathbf{G}_i^\top are the i th rows of design matrices \mathbf{B} and \mathbf{G} . Although this log-likelihood can be maximized directly, a particularly convenient method to obtain the maximum likelihood estimator (MLE) is to capitalize on the mixture structure of the problem and use the EM algorithm.

Hall (2000) extended Lambert's model and methodology to an upper bounded count situation, thereby obtaining a zero-inflated binomial (ZIB) regression model. Another popular model is the zero-inflated negative binomial (ZINB) regression model, which is defined similarly by swapping the negative binomial distribution instead of the Poisson component in a ZIP model. Although the focus of this paper is to develop robust estimation for ZIP regression models, the methods can be extended to other ZI models in the same manner.

In mixture models and more generally, the robustness of the MLE has been studied extensively. It is well known that the MLE can be unstable when the data have contamination points. In this paper, we propose an alternative approach, which we term robust expectation-solution (RES) estimation, and which is related to M-estimation. Huber (1964) proposed M-estimation as a generalization of ML in which the score function in the likelihood equation is replaced by an estimating function, which is typically chosen to downweight the contributions of extreme observations. Recently, several authors have extended M-estimation to the GLM context for independent observations (Cantoni & Ronchetti, 2001; Adimari & Ventura, 2001) and for the clustered/longitudinal data setting (Preisser & Qaqish, 1999; Cantoni, 2004). In GLM, M-estimation typically proceeds by solving estimating equations in which the contributions of observations with large Pearson residuals are reduced by downweighting. In a mixture context, however, this approach has obvious drawbacks. For example, in a 50:50 mixture of well-separated components, all observations are far from the overall mean and will necessarily have large Pearson residuals that should not be downweighted in forming an overall estimation criterion. In this paper, we get around this problem by applying M-estimation to the M-step of an EM algorithm (or, to be precise, an expectation-solution or ES algorithm). Roughly speaking, this approach effectively imputes which component each observation comes from, and then downweights the contributions of observations that are extreme in terms of low probability relative to the component to which they belong.

Recently, Lu *et al.* (2003) proposed a minimum Hellinger distance (MHD) approach for finite mixtures of Poisson regression models, a class of models in which the ZIP model falls. These authors present simulation results that suggest that their approach performs very well relative to ML in the presence of outliers and/or poor separation between the mixture components. As we will see, however, this approach has a limited domain of application because of identifiability problems that can arise when the mixing probability depends upon covariates, which is typical in applications of ZI regression. In addition, MHD approaches are not particularly effective for mixture models when the components are well separated. Furthermore, the asymptotics of MHD estimation in the regression context are difficult to establish. Therefore, we propose the RES approach and include MHD estimation as a standard of comparison.

The organization of this paper is as follows. Section 2 introduces the RES methodology and discusses its robustness and efficiency properties. Section 3 describes MHD estimation for ZIP models and addresses the identifiability problem that arises with this approach. Simulation results are presented in section 4 to compare these two methods with ML estimation. To illustrate the methodology, section 5 discusses the modelling of data from an investigation of factors predictive of youth involvement in 'use of force' (UOF) incidents during detention in Georgia state facilities. Finally, some discussion and concluding remarks are provided in section 6.

2. The robust expectation solution approach for ZIP regression

2.1. The RES algorithm

The RES algorithm we propose is a modification of the EM algorithm with the property of robustness. In ZIP models, as in other mixture models, the EM algorithm is a particularly convenient approach for computing MLE (e.g. Lambert, 1992). This algorithm is set up by introducing ‘missing data’ into the problem. In particular, suppose we knew which zeros came from the degenerate distribution (the zero state); and which came from the non-degenerate distribution (the non-zero state). That is, suppose we could observe $z_i = 1$ when y_i is from zero state, and $z_i = 0$ when y_i is from the non-zero state. Then the log-likelihood for the complete data (\mathbf{y}, \mathbf{z}) would be

$$\begin{aligned} \ell^c(\boldsymbol{\gamma}, \boldsymbol{\beta}; \mathbf{y}, \mathbf{z}) &= \sum_{i=1}^n \{z_i \mathbf{G}_i^T \boldsymbol{\gamma} - \log(1 + e^{\mathbf{G}_i^T \boldsymbol{\gamma}})\} \\ &\quad + \sum_{i=1}^n (1 - z_i) \{y_i \mathbf{B}_i^T \boldsymbol{\beta} - e^{\mathbf{B}_i^T \boldsymbol{\beta}} - \log(y_i!)\} \\ &\equiv \ell_\gamma^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z}) + \ell_\beta^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z}), \end{aligned}$$

where $\mathbf{z} = (z_1, \dots, z_n)^T$. This log-likelihood is easy to maximize, because $\ell_\gamma^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z})$ and $\ell_\beta^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z})$ can be maximized separately with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ respectively, via standard calculations. With the EM algorithm, the log-likelihood of model (1) is maximized iteratively by alternating between estimating z_i by its conditional expectation under the current estimates of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ (E step) and then, with the z_i fixed at their expected values from the E step, maximizing $\ell^c(\boldsymbol{\beta}, \boldsymbol{\gamma}; \mathbf{y}, \mathbf{z})$ (M step), until the estimated $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ converges and iteration stops.

In more detail, the EM algorithm begins with starting values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\gamma}^{(0)T}, \boldsymbol{\beta}^{(0)T})^T$ and proceeds iteratively via the following three steps until convergence.

E step. Estimate z_i by its conditional mean $z_i^{(r)}$ under the current estimates $\boldsymbol{\gamma}^{(r)}$ and $\boldsymbol{\beta}^{(r)}$

$$\begin{aligned} z_i^{(r)} &= P(\text{zero state} | y_i, \boldsymbol{\gamma}^{(r)}, \boldsymbol{\beta}^{(r)}) \\ &= \frac{P(y_i | \text{zero state})P(\text{zero state})}{P(y_i | \text{zero state})P(\text{zero state}) + P(y_i | \text{Poisson state})P(\text{Poisson state})}. \end{aligned}$$

M step for $\boldsymbol{\gamma}$. Find $\boldsymbol{\gamma}^{(r+1)}$ by maximizing $\ell_\gamma^c(\boldsymbol{\gamma}; \mathbf{y}, \mathbf{z}^{(r)})$. This can be accomplished by fitting a binomial logistic regression of $\mathbf{z}^{(r)}$ on design matrix \mathbf{G} with binomial denominator equal to one. It is equivalent to solving the estimating equation

$$\frac{1}{n} \sum_{i=1}^n \{z_i^{(r)} - \text{logit}^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})\} \mathbf{G}_i = \mathbf{0}. \tag{2}$$

M step for $\boldsymbol{\beta}$. Find $\boldsymbol{\beta}^{(r+1)}$ by maximizing $\ell_\beta^c(\boldsymbol{\beta}; \mathbf{y}, \mathbf{z}^{(r)})$. It is equivalent to solving the estimating equation

$$\frac{1}{n} \sum_{i=1}^n (1 - z_i^{(r)}) \{y_i - e^{\mathbf{B}_i^T \boldsymbol{\beta}}\} \mathbf{B}_i = \mathbf{0}. \tag{3}$$

In the RES approach, we propose to replace the estimating functions (2) and (3) from the M step of the EM algorithm with robustified estimating functions. Thus, we change from an EM algorithm to an RES algorithm. Essentially, we propose to downweight observations that fall in the extreme upper and lower tail of the Poisson distribution in the estimating function. Specifically, we suggest that $\boldsymbol{\gamma}^{(r+1)}$ and $\boldsymbol{\beta}^{(r+1)}$ be found by solving the following equations:

$$\frac{1}{n} \sum_{i=1}^n \omega(\mathbf{G}_i) \{z_i^{(r)} - \text{logit}^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})\} \mathbf{G}_i = \mathbf{0}, \tag{4}$$

$$\frac{1}{n} \sum_{i=1}^n (1 - z_i^{(r)}) \omega(\mathbf{B}_i) \{\psi_c(y_i) - o_i(\boldsymbol{\beta}, c)\} \mathbf{B}_i = \mathbf{0}, \tag{5}$$

where

$$\psi_c(y) = \begin{cases} j_1, & y < j_1, \\ y, & y \in [j_1, j_2], \\ j_2, & y > j_2, \end{cases}$$

with j_1 and j_2 being the c and $(1 - c)$ quantiles of the non-degenerate Poisson component, respectively; and $o_i(\boldsymbol{\beta}, c) = E\{\psi_c(Y_i) | Y_i \sim \text{Poisson}(\mu_i = e^{\mathbf{B}_i \boldsymbol{\beta}})\} = j_1 P(Y_i < j_1) + \mu_i P(j_1 - 1 \leq Y_i < j_2) + j_2 P(Y_i > j_2)$, where probabilities are computed based on the Poisson component density. Here, $\omega(\cdot)$ is a function to downweight large leverage points. A simple choice for $\omega(\mathbf{G}_i)$ that we use throughout this paper is $\sqrt{1 - h_i}$, where h_i is the i th diagonal element of $\mathbf{H} = \mathbf{G}(\mathbf{G}^T \mathbf{G})^{-1} \mathbf{G}^T$, with a similar definition for $\omega(\mathbf{B}_i)$. More sophisticated choices for $\omega(\cdot)$ based on, e.g. Mahalanobis distance are available (Cantoni & Ronchetti, 2001). The choice of upper and lower quantile c , in $\psi_c(\cdot)$ controls the trade-off between robustness and efficiency. In the simulations and real data example of this paper we take $c = 0.01$, a value that has been chosen to be small to guard against the occurrence of a small number of truly anomolous (or even erroneous) observations rather than to eliminate data that come from a real, non-trivial component of the mixture (i.e. a third latent class underlying the data that arises with non-negligible probability). We investigate the effect of the choice of c on asymptotic relative efficiency in section 2.4, and also comment on results for a larger value of c in the example of section 5. Other choices for the ψ_c function were considered by Shen (2006).

It is worth noting that (4) is a Mallows’s type robust estimation equation. This type of robust method is well established in the logistic regression context (Mallows, 1975; Carroll & Pederson, 1993), which is essentially that of the latent Bernoulli variable z_i . Although we did not include a ψ_c function (i.e. Huber function) in (4), the robustness in the estimation of $\boldsymbol{\gamma}$ comes from two sources: (i) at the E step, a robustly estimated $\boldsymbol{\beta}$ leads to a more accurately imputed value $z^{(r)}$, which imparts robustness to the estimation of $\boldsymbol{\gamma}$; and (ii) the Mallows’s type estimation equation helps by downweighting high leverage observations. For a discussion of the problems associated with the inclusion of a Huber function in binary logistic regression see Copas (1988) and Carroll & Pederson (1993).

The ML EM algorithm involves iteratively fitting a GLM with a weighted version of the standard ML estimating equations for a GLM, where the weights are recomputed at each iteration in the E step. In the RES algorithm, we instead iteratively fit a GLM with a weighted version of M-estimating equations. This fitting (the S step) can be implemented via a Newton–Raphson algorithm. To facilitate convergence, it is important to start with a good initial value. For $\boldsymbol{\gamma}$, one can follow the same approach as described by Lambert (1992). For $\boldsymbol{\beta}$, either the least median squares or least trimmed squares estimators for linear regression models (Rousseeuw, 1984) can be adapted (Shen, 2006).

2.2. Influence function (IF)

The IF is a useful and popular tool for quantifying the degree of robustness of a statistic by measuring the potential effect of an additional observation. The classical ML estimating equations for $\boldsymbol{\theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\beta}^T)^T$ can be written as joint equations $\frac{1}{n} \sum_{i=1}^n \{E_{\boldsymbol{\theta}}(z_i | y_i) - \text{logit}^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})\} \times \mathbf{G}_i = \mathbf{0}$, and $\frac{1}{n} \sum_{i=1}^n \{1 - E_{\boldsymbol{\theta}}(z_i | y_i)\} \{y_i - \exp(\mathbf{B}_i^T \boldsymbol{\beta})\} \mathbf{B}_i = \mathbf{0}$, where the expectation is with respect

to z_i given $Y_i = y_i$. The IF of $\hat{\beta}_{MLE}$, the MLE with respect to β for the ZI model, quantifies the influence of one additional observation y_j drawn from model (1). This function is given by $IF_{MLE}(y_j) = \{1 - E_\theta(z_j | y_j)\} \{y_j - e^{\beta_j^T \beta}\} \{-E_\theta(\frac{\partial}{\partial \beta^T} [\{1 - E_\theta(z_j | y_j)\} \{y_j - e^{\beta_j^T \beta}\} \mathbf{B}_j])\}^{-1} \mathbf{B}_j$. As can be seen in $IF_{MLE}(y_j)$, the influence of an outlier on the MLE is proportional to the score function and is, therefore, unbounded in general. The estimating functions underlying the RES method are:

$$\frac{1}{n} \sum_{i=1}^n \omega(\mathbf{G}_i) \{E_\theta(z_i | y_i) - \text{logit}^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})\} \mathbf{G}_i = \mathbf{0}, \tag{6}$$

$$\frac{1}{n} \sum_{i=1}^n \{1 - E_\theta(z_i | y_i)\} \Psi(\boldsymbol{\beta}, y_i) = \mathbf{0}, \tag{7}$$

where $\Psi(\boldsymbol{\beta}, y_i) = \omega(\mathbf{B}_i) \{\psi_c(y_i) - o_i(\boldsymbol{\beta}, c)\} \mathbf{B}_i$. Then the IF of $\hat{\beta}_{RES}$ is $IF_{RES}(y_j) = \{1 - E_\theta(z_j | y_j)\} \times \{-E_\theta(\frac{\partial}{\partial \beta^T} [\{1 - E_\theta(z_j | y_j)\} \Psi(\boldsymbol{\beta}, y_j)])\}^{-1} \Psi(\boldsymbol{\beta}, y_j)$. The IF of RES estimator is bounded because the estimating function Ψ is bounded. Therefore, $\hat{\beta}_{RES}$ is as called as *B-robust* (Hampel *et al.*, 1986). Similarly, $\hat{\gamma}_{RES}$ is B-robust because of the boundedness of the estimating function (6).

2.3. Asymptotics

For simplicity, we combine (6) and (7) and rewrite them as one equation:

$$U(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n E_\theta \{s_i(y_i, z_i, \boldsymbol{\theta}) | y_i\} = \mathbf{0},$$

with the expectation taken with respect to z_i given y_i . Here, $s_i(y_i, z_i, \boldsymbol{\theta}) = (s_{i1}(y_i, z_i, \boldsymbol{\theta})^T, s_{i2}(y_i, z_i, \boldsymbol{\theta})^T)^T$, with $s_{i1}(y_i, z_i, \boldsymbol{\theta}) = \omega(\mathbf{G}_i) \{z_i - \text{logit}^{-1}(\mathbf{G}_i^T \boldsymbol{\gamma})\} \mathbf{G}_i$, and $s_{i2}(y_i, z_i, \boldsymbol{\theta}) = (1 - z_i) \Psi(\boldsymbol{\beta}, y_i)$. Rosen *et al.* (2000) show that under certain regularity conditions, if there exists a point $\hat{\boldsymbol{\theta}}$ such that $\lim_{r \rightarrow \infty} \boldsymbol{\theta}^{(r)} = \hat{\boldsymbol{\theta}}$ where $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots$ is a sequence generated by the ES algorithm, then $\hat{\boldsymbol{\theta}}$ satisfies: (i) $U(\hat{\boldsymbol{\theta}}; \mathbf{y}) = \mathbf{0}$, and (ii) $U(\boldsymbol{\theta}; \mathbf{y})$ is an unbiased estimating function, satisfying $E_\theta \{U(\boldsymbol{\theta}; \mathbf{y})\} = \mathbf{0}$ for all $\boldsymbol{\theta}$.

The conditions of this theory are easily verified for the proposed RES algorithm applied to ZIP regression. Therefore, if the RES algorithm converges, it converges to a solution $\hat{\boldsymbol{\theta}}$ of an unbiased estimating equation. Moreover under mild regularity conditions (e.g. Carroll *et al.*, 1995, §A.3), the RES estimator $\hat{\boldsymbol{\theta}} = (\hat{\gamma}^T, \hat{\beta}^T)^T$ is consistent: $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}$, almost surely; and asymptotically normal: $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \mathbf{V})$. Here, $\mathbf{V} = \mathbf{M}^{-1} \mathbf{Q} \mathbf{M}^{-1}$ with $\mathbf{Q} = E\{n \mathbf{U}(\boldsymbol{\theta}; \mathbf{y}) \mathbf{U}(\boldsymbol{\theta}; \mathbf{y})^T\}$ and $\mathbf{M} = -E\{\dot{\mathbf{U}}(\boldsymbol{\theta}; \mathbf{y})\}$, where $\dot{\mathbf{U}} = \partial \mathbf{U} / \partial \boldsymbol{\theta}^T$. The asymptotic variance of $\hat{\boldsymbol{\theta}}$ can be estimate by $\mathbf{V}_n = \mathbf{M}_n^{-1} \mathbf{Q}_n \mathbf{M}_n^{-1}$ at $\hat{\boldsymbol{\theta}}$, where $\mathbf{M}_n = -\frac{1}{n} \sum_{i=1}^n E_\theta \{s_i(y_i, z_i, \boldsymbol{\theta}) | y_i\}$, and $\mathbf{Q}_n = \frac{1}{n} \sum_{i=1}^n [E_\theta \{s_i(y_i, z_i, \boldsymbol{\theta}) | y_i\} \times E_\theta \{s_i(y_i, z_i, \boldsymbol{\theta}) | y_i\}^T]$. However as pointed out in Rosen *et al.* (2000), this algorithm is no longer guaranteed to converge because of the fact that $s_i(y_i, z_i, \boldsymbol{\theta})$ is no longer a score function. Although in practice we have never had difficulty with the RES algorithm failing to converge, it is always prudent to attend to the progress of the individual iterates and, in some cases, starting the algorithm from multiple initial values may be useful.

2.4. Asymptotic efficiency relative to the MLE

The robustification of the estimating functions in (6) and (7) can certainly be expected to result in some loss of efficiency relative to the MLE. Here we investigate this efficiency loss and its dependence upon the tuning quantile c through a small study of the RES and ML estimators under relatively simple models from which we can compute asymptotic variances. In

Table 1. Relative efficiencies for zero-inflated Poisson (ZIP) model parameters for different values of the tuning quantile c

c	ZIP with $\gamma=1.0$				ZIP with $\gamma=4.0$			
	$\gamma=1.0$	$\beta_1=0$	$\beta_2=1$	$\beta_3=0.1$	$\gamma=4.0$	$\beta_1=0$	$\beta_2=1$	$\beta_3=0.1$
$c=0.005$	0.998	0.993	0.996	0.996	1.000	0.995	0.996	0.996
$c=0.010$	0.997	0.987	0.993	0.993	1.000	0.990	0.993	0.993
$c=0.015$	0.996	0.985	0.990	0.990	1.000	0.988	0.990	0.990
$c=0.020$	0.995	0.982	0.988	0.988	1.000	0.986	0.988	0.988
$c=0.025$	0.995	0.981	0.986	0.986	1.000	0.985	0.986	0.986
$c=0.030$	0.992	0.967	0.981	0.981	0.999	0.975	0.982	0.981
$c=0.040$	0.990	0.955	0.972	0.971	0.999	0.965	0.973	0.972
$c=0.050$	0.988	0.947	0.969	0.968	0.999	0.959	0.969	0.969
$c=0.075$	0.985	0.930	0.946	0.945	0.999	0.944	0.947	0.946
$c=0.100$	0.980	0.924	0.938	0.937	0.998	0.940	0.940	0.939

particular, we consider ZIP models with non-constant mixing probability $p = \{1 + \exp(-\gamma g)\}^{-1}$ and Poisson mean $\mu = e^{B\beta}$. Specifically, we set $g = \mathbf{j}_{10}^T \otimes (-1.5, -1.4, -1.3, -1.2, -1.1, -1.0, -0.9, -0.8, -0.7, -0.6)$ and

$$B^T = \mathbf{j}_{10}^T \otimes \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}.$$

Parameters γ and β were chosen to yield some overlap between the mixture components and are listed in Table 1. Two values, 1 and 4, were chosen for γ , which give values of p in the range (0.18, 0.35) and (0.0025, 0.083), respectively.

Based on these models, we report the ratios of asymptotic variances for values of c ranging from 0.005 to 0.10 in Table 1. Because no high leverage points exist in the data, the ratios of asymptotic variance corresponding to the mixing probability are close to 1. Ratios for the regression parameters decrease as expected with c , but the relative efficiency for the RES estimators is reasonably high even for $c=0.10$. In the simulation studies and example of sections 4 and 5, we take $c=0.01$ for the ZIP models we consider. In the simple models we consider here, this choice leads to relative efficiencies for β of 98.7 per cent or better. Of course, the relative efficiency of the RES approach will differ across models, but the results we present here give some sense of the efficiency loss that one might expect in practice.

3. MHD estimation in ZI regression models

3.1. MHD method

The literature on MHD estimation concentrates primarily on the i.i.d. setting. In that context the Hellinger distance is defined as $H^2(f_n, f_\theta) = \int \{f_n(y)^{1/2} - f_\theta(y)^{1/2}\}^2 dy$ where f_θ is the marginal density of a response variable Y based on a sample y_1, \dots, y_n according to the parametric model under consideration and f_n is a corresponding non-parametric density estimate. For discrete data, the natural choice for f_n is the empirical frequency function defined by

$$f_n(y) = N_y/n, \quad y \in \mathcal{Y}, \tag{8}$$

where N_y is the number of observations having value y , and \mathcal{Y} is the sample space for Y .

In the context of Poisson mixture regression models, Lu *et al.* (2003) presented MHD estimators defined in terms of the classical Hellinger distance between the empirical frequency function $f_n(\cdot)$ and the marginal density for Y implied by the model. Because ZIP regression falls in the class of models these authors considered, here we treat their approach as an

alternative robust estimation method and standard of comparison for the RES method of section 2. Lu *et al.* (2003) suggested replacing $f_\theta(y)$ with a simple, but consistent estimator $f_{\theta,n}(y)$, defined by

$$f_{\theta,n}(y) = \frac{1}{n} \sum_{i=1}^n f_\theta(y | \mathbf{x}_i), \quad y \in \mathcal{Y}. \tag{9}$$

This leads to an MHD estimator defined as $\hat{\theta}_{\text{MHD}} = \arg \min_{\theta \in \Theta} H^2(f_{\theta,n}, f_n)$.

Lu *et al.* (2003) claim that the results of the earlier authors should extend to the finite mixture of Poisson regression context, of which the ZIP model is a special case. These authors propose an asymptotic variance estimator of the form $\text{var}(\hat{\theta}) = \sum_{y \in \mathcal{Y}} \{ \hat{l}_\theta(y) \hat{l}_\theta^T(y) f_n(y) - \frac{\partial^2 f_{\theta,n}(y)}{\partial \theta \partial \theta^T} |_{\theta = \hat{\theta}} \}$, where $\hat{l}_\theta(y) = \frac{\partial}{\partial \theta} \log f_{\theta,n}(y) |_{\theta = \hat{\theta}}$. However, formal proofs for asymptotic properties of MHD estimators in the non-i.i.d. set-up have not been established.

3.2. Identifiability

The issue of identifiability of finite mixture models has attracted considerable attention in the literature, but most of this discussion has centred on the likelihood function and has assumed constant mixing probability in the model. When using MHD estimation via marginal densities rather than ML, however, the class of identifiable models is more restricted. In addition, ZI regression models allow a regression structure $\text{logit}(p_i) = \mathbf{G}_i^T \boldsymbol{\gamma}$, which invalidates many of the existing results and adds complexity to the identifiability question. For ZI models with non-constant mixing probability it is not hard to find simple non-identifiable models based on the unconditional (marginal) density. For example, let

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{Poisson}(\mu), & \text{with probability } 1 - p_i, \end{cases}$$

and suppose that p_i depends on X_i , where X_i is a binary variable,

$$X_i = \begin{cases} 0, & i = 1, 2, \dots, \frac{n}{2} \\ 1, & i = \frac{n}{2} + 1, \dots, n. \end{cases}$$

It follows that $p_i = \text{logit}^{-1}(\gamma_1) \equiv P_1$ if $i = 1, 2, \dots, n/2$, otherwise $p_i = \text{logit}^{-1}(\gamma_2) \equiv P_2$. The marginal density is $f_{\theta,n}(y) = \frac{1}{2}(P_1 + P_2)I(y=0) + \{1 - \frac{1}{2}(P_1 + P_2)\} \frac{e^{-\mu} \mu^y}{y!}$, which is clearly not identifiable.

Necessary and sufficient identifiability conditions in the class of models are difficult to establish. However, in our simulation studies we encountered singular Hessian matrices for the MHD criterion for most of the models we considered that have non-constant mixing probability.

4. Simulation studies

The aim of these simulations is to assess the performance of ML, RES and MHD estimators in the presence of outliers and/or poor separation of the mixture components. Because of non-identifiability problems with the MHD method, we restrict attention to the case of constant mixing probability in study 1, and consider only the RES and ML approaches for non-constant mixing probability in simulation studies 2 and 3. In studies 1 and 2, data are generated from a ZIP model with outliers in the response, whereas in study 3 we generate data from a ZIP model and add outlying values in the explanatory variables (high leverage points). In all the three studies, two sample sizes, $n = 100$ and $n = 200$, and two different degrees of separation between the mixture components are examined.

All simulations involve data generated from a model of the form

$$Y_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{Poisson}(\mu_i), & \text{with probability } 1 - p_i, \end{cases}$$

where $\mu_i = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i})$, and

$$p_i = \begin{cases} p, & \text{in simulation study 1,} \\ \text{logit}^{-1}(\gamma_1 x_{1i} + \gamma_2 x_{2i} + \gamma_3 x_{3i}), & \text{in simulation studies 2 and 3.} \end{cases}$$

Covariates here are $x_{1i} = I(i \leq \frac{n}{2})$, $x_{2i} = I(i > \frac{n}{2})$, $x_{3i} \sim U(0, 1)$ and $x_{4i} \sim N(0, 1)$. The data were generated under various settings of the model parameters γ and β , chosen to correspond to low versus high levels of ZI combined with low versus high levels of separation between the mixture components. For every parameter setting, 500 data sets were generated, with outliers added depending on the data contamination scheme under consideration. Bias, mean square error (MSE) and empirical size of a nominally 0.05-level Wald test for equality with the true value were calculated for each model parameter. In addition, we provided the MSE for $\zeta \equiv \frac{1}{n} \sum_{i=1}^n (1 - p_i) \mu_i$, the average marginal mean according to the model. In all three simulation studies the tuning quantile c was set to 0.01.

4.1. Study 1: ZIP regression with constant p and outliers in y

In study 1, we compare ML, MHD and RES estimation methods for ZIP data with constant p with and without contamination in y . In the contaminated scenario, 5 per cent of the response y in each data set were randomly selected to be replaced by $y + 15$. True values of p and β were specified as listed in Table 2 and were chosen to make the non-degenerate component's mean large (μ ranging between 2.78 and 20 over the values of the covariate vector $\mathbf{x}_i = (x_{1i}, \dots, x_{4i})^T$) and to give a moderate level of ZI (20 per cent). Results appear in Table 2.

Generally speaking, these results favour the RES approach over both the ML and MHD estimations. In the absence of contamination RES performs slightly worse than ML for $n = 100$ and essentially the same for the larger sample size. The MHD approach is not competitive in this setting. When contamination was present, there are a few cases for which the RES estimators exhibited greater bias than those of the MHD approach, but RES bias was generally lower than that of ML and, with few exceptions, the MSE was much smaller for RES than either MHD or ML estimation. In addition, the size of the Wald tests was much more severely altered by the presence of outliers under the ML estimation than the MHD and RES approaches. It should be kept in mind that the degree of contamination here is fairly extreme. Both the proportion (5 per cent) and magnitude ($y + 15$) of outliers here are quite large. Under these extreme circumstances, the Wald tests under RES and MHD estimation perform reasonably well, and seem to retain some value as inferential tools. In contrast, the tests under ML estimation have been completely undermined.

To examine the effect of more moderate degrees of contamination, we ran simulations similar in design to these but with 5 per cent of the responses increased by 7 rather than 15. The results from those simulations are similar to those from the bottom half of Table 2, with smaller but still quite substantial improvements in bias and MSE achieved using the RES method. Because these results are, as one might expect, intermediate to those in the top and bottom halves of Table 2, we do not report them in detail here for the sake of brevity.

The performance of MHD estimation relative to ML is somewhat surprising especially for $n = 100$. In simulation results covering the case in which the mixture components are poorly separated (not reported here but see Shen, 2006), we have observed reductions in MSE for MHD relative to ML comparable with the gains exhibited by RES estimation. In this

Table 2. Simulation results for a zero-inflated Poisson model with well-separated components, constant mixing probability $p=0.2$, and 0 per cent (top half of table) or 5 per cent (bottom half of table) outliers in y

Parameters	RES			MHDE			MLE		
	Bias	MSE	Size	Bias	MSE	Size	Bias	MSE	Size
<i>n</i> = 100									
$p=0.2$	-0.0001	0.0016	0.04	0.0177	0.0023	0.07	0.0003	0.0016	0.04
$\beta_1=1.0$	-0.0111	0.0143	0.08	0.0701	0.0469	0.07	-0.0106	0.0141	0.07
$\beta_2=2.0$	-0.0060	0.0090	0.07	0.00007	0.0510	0.05	-0.0058	0.0090	0.05
$\beta_3=1.0$	0.0024	0.0220	0.07	-0.1832	0.1399	0.09	0.0017	0.0219	0.06
$\beta_4=0.1$	-0.0018	0.0017	0.08	-0.0279	0.0019	0.00	-0.0016	0.0017	0.06
ζ		0.5222			2.0016			0.5210	
<i>n</i> = 200									
$p=0.2$	-0.0011	0.0008	0.05	0.0076	0.0009	0.05	-0.0164	0.0008	0.05
$\beta_1=1.0$	-0.0092	0.0066	0.06	0.0435	0.0148	0.02	-0.0094	0.0066	0.06
$\beta_2=2.0$	-0.0015	0.0040	0.05	0.0227	0.0159	0.02	-0.0016	0.0040	0.04
$\beta_3=1.0$	0.0049	0.0087	0.04	-0.1239	0.0463	0.05	0.0048	0.0087	0.04
$\beta_4=0.1$	-0.0006	0.0008	0.07	-0.0305	0.0024	0.00	-0.0006	0.0008	0.05
ζ		0.2628			0.8396			0.2628	
<i>n</i> = 100									
$p=0.2$	-0.0075	0.0016	0.05	0.0082	0.0020	0.07	-0.0055	0.0016	0.07
$\beta_1=1.0$	0.1745	0.0431	0.18	0.0663	0.0468	0.09	0.3326	0.1211	0.88
$\beta_2=2.0$	0.1050	0.0193	0.17	0.0859	0.0512	0.09	0.1731	0.0372	0.55
$\beta_3=1.0$	-0.1434	0.0412	0.12	-0.1358	0.1069	0.07	-0.2332	0.0731	0.43
$\beta_4=0.1$	0.0097	0.0015	0.07	-0.0253	0.0015	0.00	0.0046	0.0016	0.07
ζ		0.7743			1.0236			1.2133	
<i>n</i> = 200									
$p=0.2$	-0.0091	0.0008	0.04	-0.0012	0.0008	0.04	-0.0069	0.0008	0.06
$\beta_1=1.0$	0.1152	0.0194	0.18	0.0470	0.0157	0.04	0.2706	0.0779	0.96
$\beta_2=2.0$	0.0833	0.0107	0.19	0.0708	0.0205	0.05	0.1585	0.0283	0.74
$\beta_3=1.0$	-0.0717	0.0133	0.08	-0.0702	0.0381	0.05	-0.1660	0.0347	0.44
$\beta_4=0.1$	-0.0062	0.0008	0.06	-0.0145	0.0022	0.00	-0.0171	0.0010	0.09
ζ		0.4939			0.6196			0.8817	

RES, robust expectation-solution; MHDE, minimum Hellinger distance estimation; MLE, maximum likelihood estimation; MSE, mean square error.

study, however, MHD estimation does not demonstrate a consistent advantage over ML with respect to β . Increasing sample size from $n = 100$ to 200 has the expected effect of decreasing MSE for all parameters across all the three methods.

4.2. Study 2: ZIP regression with non-constant p and outliers in y

Here, data were generated from a ZIP regression with non-constant p in the same way as described in study 1. True values for the parameters γ and β are specified in Table 3. These parameter values correspond to a relatively large proportion of zeros (p ranging from 0.27 to 0.37) and poor separation between the components (μ ranging from 1 to 7).

Results for this case (see Table 3) echo those of study 1. Again, RES demonstrates substantially less bias and MSE than ML. These advantages are greatest for β and ζ , but are also present for γ . RES also performs much better in terms of Wald test size.

4.3. Study 3: ZIP regression with non-constant p and outliers in x

Here we generated data from well-separated ZIP models with relatively large proportion of zeros (p ranging from 0.12 to 0.27 and μ ranging from 1.00 to 110 with average 15). To create outliers in x , we randomly chose about 1 per cent of the observations (one point for

Table 3. *Simulation results for a zero-inflated Poisson model with poorly separated components, regression structure for the mixing probability p , and 5 per cent outliers in y*

Parameters	RES			MLE		
	Bias	MSE	Size	Bias	MSE	Size
<i>n</i> = 100						
$\gamma_1 = -1$	0.1385	0.4150	0.05	0.5361	0.5286	0.20
$\gamma_2 = -0.5$	0.0185	0.3155	0.05	0.1764	0.3142	0.08
$\gamma_3 = 0.5$	-0.2276	0.7497	0.06	-0.4988	0.7875	0.10
$\beta_1 = 0.0$	0.3125	0.1626	0.23	0.8265	0.7401	0.92
$\beta_2 = 1.0$	0.1582	0.0737	0.20	0.3969	0.2115	0.62
$\beta_3 = 1.0$	-0.1073	0.1294	0.10	-0.3819	0.3104	0.48
$\beta_4 = 0.1$	-0.0086	0.0145	0.17	-0.0386	0.0189	0.40
ζ		0.5391			1.1288	
<i>n</i> = 200						
$\gamma_1 = -1$	0.1771	0.2318	0.10	0.5616	0.4547	0.38
$\gamma_2 = -0.5$	0.0133	0.1360	0.07	0.1558	0.1411	0.09
$\gamma_3 = 0.5$	-0.1663	0.3609	0.07	-0.4171	0.4487	0.13
$\beta_1 = 0.0$	0.2802	0.1232	0.28	0.8108	0.7236	0.97
$\beta_2 = 1.0$	0.1699	0.0548	0.27	0.4179	0.2132	0.83
$\beta_3 = 1.0$	-0.0876	0.0804	0.15	-0.3669	0.2538	0.59
$\beta_4 = 0.1$	-0.0134	0.0048	0.10	-0.0434	0.0088	0.34
ζ		0.3001			0.8284	

RES, robust expectation-solution; MLE, maximum likelihood estimation; MSE, mean square error.

$n=100$ and two points for $n=200$), and replaced the covariate value x_3 by $x_3 + 3$, leaving the response y and other covariates unchanged.

The results in Table 4 show that RES and ML perform similarly with respect to γ . This result is sensible, as there is only a small amount of contamination in \mathbf{x} that is of a form that does not obscure the mixture structure underlying the data. With respect to β and ζ , however, RES has less bias, smaller MSE and closer to nominal size than ML estimation. Generally speaking, the usual positive effect of sample size on efficiency is observed.

5. Example

To illustrate the use of robust methods for ZI regression models, we consider data collected by Jackson (2004) who investigated factors predictive of youth involvement in adverse incidents during residence in Georgia state detention facilities. Data on $n=13,517$ youths detained in the state of Georgia during the period 1 July 2001 to 30 June 2002 were collected from the Juvenile Tracking System, a database maintained by the Georgia Department of Juvenile Justice. These data contain information on detainees' involvement in certain types of incidents (e.g. allegations of child abuse, suicide attempts, youth on youth assault, etc.) as well as characteristics of the child, the facility in which the incident occurred and the nature of the detention. Here, we consider models for detainees' involvement in one class of incidents studied by Jackson: UOF occurrences. Although some subjects were detained in multiple facilities over the period covered in the data set, for simplicity we restrict attention to each subject's first stay in a detention facility during the period of interest.

Table 5 contains a frequency distribution for the UOF counts, 94.07 per cent of which are 0, and 99.9 per cent of which are ≤ 5 . In addition, there are a few especially large counts, including one of 22 that seems particularly extreme. We consider regression models for these data built from a fairly rich set of covariates, so the marginal distribution of Table 5 gives

Table 4. Simulation results for a zero-inflated Poisson model with well-separated components, regression structure for the mixing probability p , and 1 per cent outliers in the covariates

Parameters	RES			MLE		
	Bias	MSE	Size	Bias	MSE	Size
<i>n</i> = 100						
$\gamma_1 = -2$	0.0157	0.5813	0.07	-0.0072	0.5954	0.07
$\gamma_2 = -1.5$	0.1206	0.2977	0.04	0.1446	0.3010	0.03
$\gamma_3 = 0.5$	-0.2294	0.8185	0.04	-0.2747	0.8354	0.05
$\beta_1 = 1.0$	0.0556	0.0121	0.012	0.0823	0.0151	0.15
$\beta_2 = 2.0$	0.1051	0.0174	0.12	0.1459	0.0272	0.54
$\beta_3 = 1.0$	-0.193	0.0517	0.15	-0.2705	0.0868	0.86
$\beta_4 = 1.0$	-0.0027	0.0020	0.05	0.0057	0.0019	0.04
ζ		2.7484			3.0402	
<i>n</i> = 200						
$\gamma_1 = -2$	-0.0538	0.3021	0.05	-0.0379	0.3085	0.05
$\gamma_2 = -1.5$	-0.0183	0.1968	0.07	-0.0054	0.2012	0.07
$\gamma_3 = 0.5$	0.0146	0.4029	0.03	-0.0249	0.4351	0.04
$\beta_1 = 1.0$	0.0056	0.0068	0.10	0.1039	0.0150	0.45
$\beta_2 = 2.0$	0.0526	0.0056	0.10	0.0913	0.0118	0.45
$\beta_3 = 1.0$	-0.1302	0.0235	0.12	-0.2323	0.0639	0.87
$\beta_4 = 1.0$	0.0216	0.0014	0.13	0.040	0.0027	0.35
ζ		2.960			3.5696	

RES, robust expectation-solution; MLE, maximum likelihood estimation; MSE, mean square error.

only a preliminary, rough indication that ZI and outliers may be complicating issues for these data.

Jackson (2004) considered binary response models for the occurrence of one or more UOF incidents during detention and examined explanatory variables including characteristics of the youth: age, sex, race (white, non-white), number of prior detentions (priors) and a detention assessment instrument (dai) score, which categorizes the severity of the subject's offence history; and characteristics of the detention episode: site identifier (22 sites in all), length of stay (los), average utilization (avgutil, or average proportion of beds filled during the detention) and utilization difference (utildiff, defined as the difference between the average and maximum utilization during the detention period). Here, we examine ZIP regression models for the UOF count. Model building in a ZIP regression context with a large number of covariates is a challenging task, especially under the additional complication posed by the presence of outliers. For simplicity, we avoid this process and simply consider the performance of the RES ($c=0.01$) and ML approaches to ZIP regression under a reasonable 'full model' for the use of force data. Because constant mixing probability models are overly simplistic in this example, MHD estimation is excluded from consideration.

In particular, we fit models in which the two linear predictors of the model are identically specified to include main effects of all categorical predictors (site, sex, race, dai) and linear and quadratic terms for each of the continuous predictors (age, priors, los, avgutil and utildiff). Results from fitting this model appear in Table 6, which presents parameter estimates and Wald p -values for all parameters related to the continuous predictors and Wald p -values for main effects of the categorical predictors. It is clear from Table 6 that there are substantial differences between the ML and RES parameter estimates and between the qualitative conclusions to be drawn regarding specific predictors based on the two methodologies. For instance, the statistical significance of race, dai, age and age² are all overestimated using ML, which gives p -values less than 0.05 for all of these predictors.

Table 5. Frequency distribution for use of force counts

Value	Frequency	Per cent	Cumulative per cent
0	12,715	94.07	94.07
1	573	4.24	98.31
2	151	1.12	99.42
3	40	0.30	99.72
4	12	0.09	99.81
5	13	0.10	99.90
6	5	0.04	99.94
7	1	0.01	99.95
8	1	0.01	99.96
9	1	0.01	99.96
10	2	0.01	99.98
12	1	0.01	99.99
14	1	0.01	99.99
22	1	0.01	100.00

Table 6. Parameter estimates and Wald *p*-values (in parentheses) for zero inflated Poisson model for use of force counts

Variable	Poisson component		Mixing probability	
	ML	RES	ML	RES
site	– (<0.001)	– (<0.001)	– (0.006)	– (0.023)
sex	– (0.970)	– (0.644)	– (0.014)	– (0.135)
race	– (0.022)	– (0.303)	– (0.043)	– (0.283)
dai	– (<0.001)	– (0.076)	– (0.001)	– (0.110)
age	2.16 (0.010)	1.46 (0.261)	3.27 (0.047)	2.39 (0.328)
age ²	–2.64 (0.002)	–1.91 (0.144)	–3.53 (0.036)	–2.67 (0.283)
priors	0.328 (<0.001)	0.409 (<0.001)	–0.417 (0.009)	–0.293 (0.137)
priors ²	–0.0879 (0.188)	–0.141 (0.061)	0.240 (0.149)	0.138 (0.406)
los	0.248 (0.002)	0.201 (0.038)	–1.42 (<0.001)	–1.88 (<0.001)
los ²	–0.0814 (0.051)	–0.0442 (0.392)	0.558 (<0.001)	0.741 (<0.001)
avgutil	–0.370 (0.323)	–0.281 (0.659)	0.332 (0.600)	0.491 (0.682)
avgutil ²	0.175 (0.632)	0.0694 (0.915)	–0.472 (0.445)	–0.699 (0.573)
utildiff	–0.0177 (0.908)	–0.224 (0.294)	–1.52 (<0.001)	–1.72 (<0.001)
utildiff ²	0.122 (0.172)	0.226 (0.061)	1.18 (<0.001)	1.36 (<0.001)

ML, maximum likelihood; RES, robust expectation-solution.

The significant predictors identified in the RES-fitted model largely reproduce those found by Jackson (2004). In particular, the variable priors is positively associated with the Poisson UOF incident rate as is los, which, as should be expected, has a positive influence on the Poisson rate of incidents and a significant negative effect on the probability of the zero state. utildiff, which is a measure of the variability in crowding at the detention facility, was also found to have a positive influence on the probability of being involved in UOF incidents.

The RES results in Table 6 were obtained using tuning constant $c=0.01$. In addition, we refit this model using the RES method with $c=0.025$ to investigate sensitivity to this choice. Generally speaking, effects of using this larger tuning constant were negligible on parameter estimates and *p*-values associated with γ . For β , the larger *c* value had slightly greater effects, although mainly on the Wald *p*-values rather than the parameter estimates. Because the larger *c* value effectively reduces the estimated variability in the data, *p*-values tended to drop somewhat, although not drastically; qualitative conclusions about the effects of the explanatory variables remain mostly the same, with the only change being to priors² and utildiff² whose *p*-values dropped just below the traditional 0.05 significance level.

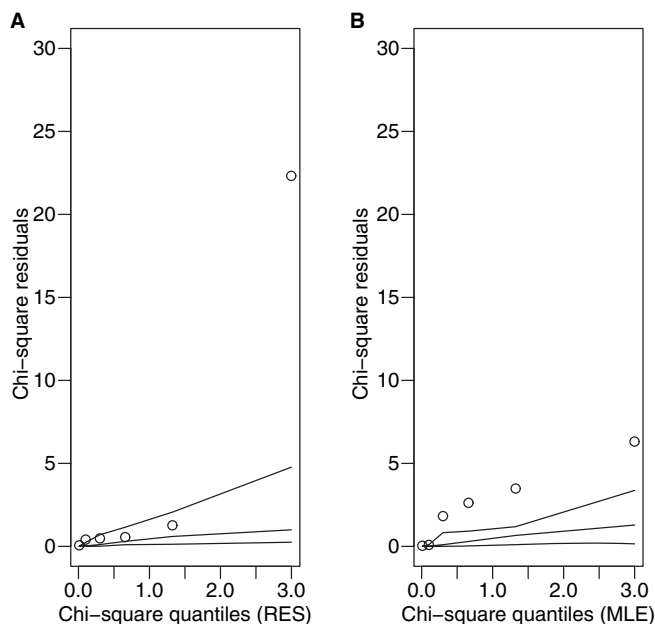


Fig. 1. Chi-square residual plots for model fit to the use of force data: (A) chi-square quantiles (robust expectation solution); (B) chi-square quantiles (maximum likelihood estimation).

To further compare the RES and ML fits of this model, we produced plots of chi-square residuals. These plots are similar in concept to half normal plots (Atkinson, 1981; Vieira *et al.*, 2000), and were constructed by first binning the data into b bins according to the observed values of the response. In this case, we used $b=6$ bins corresponding to 0, 1, 2, 3, 4, $[5, \infty)$. Then we calculated residuals defined as the contribution to the chi-square goodness-of-fit statistic $r_j = n\{f_n(y) - f_{\theta,n}(y)\}^2 / f_{\theta,n}(y)$, $j = 1, \dots, 6$, for each bin based on the fitted model (using RES or ML). The idea behind this plot is if the model is correctly specified, any collection of $b-1$ residuals should be approximately i.i.d. $\chi^2(1)$ random variables. The residual plot is a plot of ordered $r_{(j)}$ s against $\chi^2(1)$ quantiles. A simulated envelope for this chi-square plot was constructed in the same way as is typically done in half normal plots (e.g. Vieira *et al.*, 2000). The plots show that most of the residuals fall within the boundaries of the envelope using RES (Fig. 1A), but not with ML (Fig. 1B), indicating that the former method is more appropriate for these data. The one large point in Fig. 1(A) is a desirable feature. It corresponds to the outlying values in the largest bin, which should have large residuals if the estimation downweights these outliers as intended. Although our choice of binning scheme is somewhat arbitrary, we examined chi-square residual plots for other choices of the bins and obtained similar results (omitted for the sake of brevity).

Figure 2 displays the downweighting scheme for the UOF model using the RES method. In particular, we plot the quantity $\omega(\mathbf{B}_i)\{\psi_c(y_i) - o_i(\boldsymbol{\beta}, c)\} / (y_i - \mu_i)$ versus y_i , $i = 1, \dots, n$. While most observations receive weights near 1, Fig. 2 shows the expected behaviour, with downweighting tending to occur for the larger values of y , and the most extreme downweighting for the one observation for which $y = 22$.

Finally, note that a negative binomial regression model does not fit these data as well as the ZIP model, even when fit with ML estimation. For example, Akaike information criterion values (smaller is better form) for these two models are 5929.3 (negative binomial) and 5908.4 (ZIP). In addition, the negative binomial model has the disadvantage of not

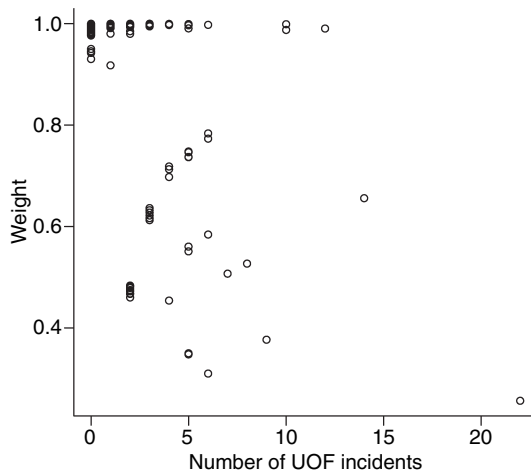


Fig. 2. Observation weights assigned by the robust expectation-solution for β in the model considered for use of force counts.

providing information about predictors of a child being at risk for becoming involved in any UOF incidents separately from predictors of the mean number of incidents among those at risk.

6. Discussion

In this paper, we proposed the RES algorithm, a novel method of robust estimation for ZIP regression. In addition, we studied the existing approach, MHD estimation, in this context, and compared these two approaches with ML. Our simulation results suggest that the RES approach provides substantial protection against outliers relative to ML and easily outperforms MHD estimation. In addition, the MHD method leads to identifiability problems for some models that are identifiable when fit with ML or the RES approach and is therefore, substantially more narrowly applicable. The estimating equations we proposed in the RES method perform well in downweighting outliers in y and/or covariates \mathbf{x} . However, our approach does require specification of the tuning quantile c , which affects the efficiency of the parameter estimators. We have provided evidence of mild losses of efficiency and good robustness properties for values of c in the range $[0.01, 0.05]$. Further research is ongoing to develop methodology for data-dependent selection of c . Recent work on this topic by Wang *et al.* (2007) seems promising and may be adaptable to our ZIP regression setting.

It is worth noting that the RES approach can be extended to the general ZI regression context easily (e.g. the ZIB model). Another natural extension of the RES approach would be to generalize this method to the clustered data context. We are currently pursuing this goal by combining our approach with that of Hall & Zhang (2004) who proposed an estimation method for marginal ZI regression models for clustered data via generalized estimating equations. This extension will lead to methodology to handle longitudinal data sets subject to ZI and extreme values.

The overall approach adopted in this paper is to accommodate, rather than eliminate, outliers and use a robust estimation methodology that minimizes their effect on estimation of the model that is followed by the vast majority of the data. In some instances, extreme values may be truly erroneous, in which case they are contaminating values of no research interest and minimization of these values' effects would seem to be uncontroversial. In other cases,

the extreme values may be legitimate and of some inherent scientific interest. For example, in the UOF data of section 5, there may be a very small proportion of juvenile detention episodes that generate very large UOF incident counts, and the circumstances that lead to these high counts (whether they be characteristics of the child or the conditions of detention) would certainly be of some interest. That is, as an anonymous referee has suggested, it may be desirable to model these extremes rather than to minimize their effect on the model. However, by their inherent rarity, modelling such values is a challenge.

Two approaches for incorporating the few extremely large values into the model without downweighting their contributions are: (i) to consider a heavier-tailed distribution such as the ZINB, and (ii) to regard these values as being generated from an additional latent class (or classes) with a higher mean and introduce one (or more) components to the mixture model (e.g. a ZI two-component Poisson model). Both of these approaches have drawbacks.

ZINB regression models have become popular tools for analysing highly dispersed count data with many zeros, and they are certainly useful. However, such models can be strongly affected by extremes as well. To illustrate this, we used the ML estimation to fit a ZINB version of the model considered in section 5 to the UOF data both with and without the one observation in the data set for which $y=22$ (the largest response). Inferences regarding covariate effects on the mixing probability p differed markedly between the ZINB models with and without this subject's data. In particular, several predictors were found to be statistically significant when this observation was omitted (los, los², priors, utildiff² all significant at $\alpha=0.05$, and priors² and age² at $\alpha=0.10$). However, when the observation was included, all p -values for covariates in $G\gamma$ increased beyond 0.20 and no covariates were found to be significant predictors of the zero state. The intuition for this seems clear: accommodation of the extreme at $y=22$ inflates the negative binomial variance function, shifting many of the zero counts from the zero state to the negative binomial component of the mixture and drastically altering the fitted model for p . Clearly, this degree of sensitivity to one subject's data (out of 13,517 total observations) is undesirable because, regardless of whether the extreme value is of interest and 'belongs' in the analysis or is a contaminant, its presence alters the conclusions to be drawn about the population from which the bulk of the data are drawn.

Alternatively, one might introduce an additional component to the mixture model to capture extreme values. This would have the advantage of retaining the ZIP structure for the majority of the data, but there seems little hope of being able to build any non-trivial model for the probability of being in the third, largest-mean component as this component would necessarily be associated with very few observations. It would seem more fruitful to identify the few outliers and examine each of these subjects' data individually to make tentative conjectures about the important predictors of large frequency counts, rather than rely on a model for these extremes. Of course, the effective identification of outliers requires a robust estimation methodology, such as the RES approach we have advocated.

Acknowledgements

The authors express their sincere gratitude to Douglas K. Jackson for providing the data analysed in section 5.

References

- Adimari, G. & Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statist. Probab. Lett.* **55**, 413–419.
- Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 13–20.

- Cantoni, E. (2004). A robust approach to longitudinal data analysis. *Canad. J. Statist.* **32**, 169–180.
- Cantoni, E. & Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96**, 1022–1030.
- Carroll, R. J. & Pederson, S. (1993). On robustness in the logistic regression model. *J. Roy. Statist. Soc. Ser. B* **55**, 693–706.
- Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement error in nonlinear models*. Chapman and Hall, New York.
- Copas, J. B. (1988). Binary regression models for contaminated data (with discussion). *J. Roy. Statist. Soc. Ser. B* **50**, 225–265.
- Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030–1039.
- Hall, D. B. & Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statist. Model.* **4**, 161–180.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics*. John Wiley and Sons, New York.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.
- Jackson, D. K. (2004). Factors that are predictive of involvement of detained youth in adverse incidents. *Unpublished dissertation*, School of Social Work, University of Georgia.
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.
- Lu, Z., Hui, Y. & Lee, A. (2003). Minimum Hellinger distance estimation for finite mixtures of Poisson regression models and its applications. *Biometrics* **59**, 1016–1026.
- Mallows, C. L. (1975). On some topics in robustness. *Technical Memorandum*. Bell Telephone Laboratories, Murray Hill.
- Preisner, J. S. & Qaqish, B. F. (1999). Robust regression for clustered data with application to binary responses. *Biometrics* **55**, 574–579.
- Rosen, O., Jiang, W. X. & Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika* **87**, 391–404.
- Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79**, 871–880.
- Shen, J. (2006). Robust estimation and inference in finite mixtures of generalized linear models. *Unpublished dissertation*, The Department of Statistics, University of Georgia.
- Vieira, A. M. C., Hinde, J. P. & Demétrio, C. G. B. (2000). Zero-inflated proportion data models applied to a biological control assay. *J. Appl. Statist.* **27**, 373–389.
- Wang, Y.-G., Lin, X., Zhu, M. & Bai, Z. (2007). Robust estimation using the Huber function with a data-dependent tuning constant. *J. Comput. Graph. Statist.* **16**, 1–14.

Received November 2008, in final form April 2009

Daniel B. Hall, Department of Statistics, The University of Georgia, Athens, GA 30602, USA.
E-mail: dhall@stat.uga.edu