CrossMark

REGULAR ARTICLE

Regression modeling of one-inflated positive count data

Fatemeh Hassanzadeh[1] · **Iraj Kazemi**[1]

Abstract This paper extends regression modeling of positive count data to deal with excessive proportion of one counts. In particular, we propose one-inflated positive (OIP) regression models and present some of their properties. Also, the stochastic hierarchical representation of one-inflated positive poisson and negative binomial regression models are achieved. It is illustrated that the standard OIP model may be inadequate in the presence of one inflation and the lack of independence. Thus, to take into account the inherent correlation of responses, a class of two-level OIP regression models with subjects heterogeneity effects is introduced. A simulation study is conducted to highlight theoretical aspects. Results show that when one-inflation or over-dispersion in the data generating process is ignored, parameter estimates are inefficient and statistically reliable findings are missed. Finally, we analyze a real data set taken from a length of hospital stay study to illustrate the usefulness of our proposed models.

1 Introduction

Related to the structure of count data, several regression models, such as the negative binomial (NB) and its zero-inflated version (e.g., Yau et al. 2003; Garay et al.

✉ Iraj Kazemi
i.kazemi@stat.ui.ac.ir

Fatemeh Hassanzadeh
f.hassanzadeh@stat.ui.ac.ir

[1] Department of Statistics, University of Isfahan, Isfahan 81746, Iran

2011) are introduced in the literature. In practical applications, the positive poisson (PP) regression model is also being used to analyze positive, or zero-truncated, count outcomes (e.g., Matthews and Appleton 1993; Xie and Aickin 1997; Zuur et al. 2009). A useful modeling strategy in the presence of over-dispersion in zero-truncated data is addressed by Sampford (1955) and then is extended by many others, such as Gurmu (1991). Some properties of the positive negative-binomial (PNB) distribution is given by Lee et al. (2003). Other count regression models are also introduced to allow over-dispersion caused by unobserved heterogeneity and excess zeros produced by rare event occurrences. Various popular approaches to deal with these issues are often verified based on zero-inflated poisson (Lambert 1992) and zero-inflated NB models (e.g., Yau et al. 2003). Recently, Lim et al. (2014) propose a zero-inflated poisson mixture model for heterogeneous count data with excess zeros. An extensive review of the related literature on the NB model including a number of applications taken from a wide variety of disciplines is provided by Hilbe (2011).

The application of count data has been extensively discussed by many authors and variety of regression models are proposed to analyze certain real-life count data sets (e.g., Gschlößl and Czado 2008; Cordeiro et al. 2012). In our knowledge, in almost all proposed cases, there is currently a gap in the existing research literature when the count response exhibits excess frequency of ones while analyzing positive outcomes. These features have motivated the introduction of new regression models for count data. Thus, we first introduce an alternative regression model, namely the one-inflated positive poisson (OIPP), to deal with occurrence of excess ones. It is discussed that over-dispersion may be the result of excess ones or some other causes. If extra variation remains even after handling excess ones, we then introduce the one-inflated positive negative-binomial (OIPNB) regression model. This model allows the variance to be larger than the mean through an additional parameter to handle over-dispersion. In general, the proposed models are shown to be constructed by a mixed strategy such that it mixes a distribution degenerated at one with a baseline PP or a NB distribution. Hence, this paper provides a useful approach in modeling positive count outcomes focusing mainly on data that exhibits over-dispersion.

In fitting zero-truncated count models, one-inflation and the lack of independence may exist simultaneously as a consequence of the inherent correlation structure and the underlying heterogeneity. Thus, this paper introduces a two-level OIP regression model as an alternative to handle clustered observations. This extension is motivated by methodologies addressed in fitting zero-inflated models (e.g., Hall 2000; Wang et al. 2002; Hur et al. 2002).

A simulation study is conducted to illustrate the importance of modeling strategies for one-inflation in positive counts and to show the impact of models mis-specification. The simulation is designed for two OIPP and OIPNB models. The average root mean squared error (RMSE) is used to assess the overall performance of estimates, while the average model bias is used to assess the impact of correctly identifying the OIPP or the OIPNB models on statistical inference.

The data set we re-analyze in this paper is originally taken from the US national Medicare inpatient hospital database which is prepared yearly from hospital filing records. Several researchers used these count data and fitted some models mainly without any concern on the positiveness of counts. In specific, Hardin and Hilbe

(2007) and Hilbe (2011) fitted poisson and NB models. They also analyzed the data set using the zero-truncated version of poisson and NB models. In the present paper we fit the proposed OIPP and OIPNB regression models and report that these are better fitted to the data.

The rest of paper is organized as follows. Section 2 introduces the OIP distribution and highlights some of its properties. Specifically, we derive the probability mass function of the OIP distribution in general, and find its corresponding recursive formula. Specific properties of the OIPP and OIPNB distributions are provided as special cases. Section 3 introduces OIP regression models and derives associated likelihood equations. Also, the stochastic hierarchical form of the OIPP and OIPNB distributions are achieved. Section 4 proposes a two-level OIP regression model with heterogeneity effects to address fitting of correlated data. A simulation study for large samples is conducted in Sect. 5 to highlight theoretical aspects. Section 6 aims to analyze length of stay in hospital data.

2 The specification of OIP distributions

Motivation to introduce OIP distributions arises originally from the fact that a variety of positive or zero-truncated count data involves excess ones. Here, the one and subsequent counts are generated by different mechanisms.

Definition 1 Let for each subject Y, a Bernoulli trial be used to determine two data generating processes for only a one response, with probability p, and a positive distribution, with probability $1-p$, for subsequent counts. This introduces the probability mass function (pmf) of OIP random variable as

$$
\begin{aligned}
f_{OIP}(1) &= p + (1-p) f_P(1), \\
f_{OIP}(Y = y) &= (1-p) f_P(y), \qquad y = 2, 3, \cdots,
\end{aligned}
\tag{1}
$$

where $f_P(y) = P(Y = y)$ denotes the pmf of a positive count at point y.

In general, the pmf of a positive count is derived in terms of its underlying un-truncated distribution. That is, for a given un-truncated pmf, $f(y)$, we have $f_P(y) = \pi f(y)$ for $y = 1, 2, \cdots$, where $\pi = (1 - f(0))^{-1}$ with $\pi > 1$. Putting this in Eq. (1) gives the OIP pmf. The OIP poisson and OIP NB distributions are two special cases of the family of OIP distributions though the OIPNB can be thought to arise as an extension of either the PNB or the OIPP distributions.

The pmf of an OIPP random variable Y is defined by setting $f_P(\cdot)$ in Eq. (1) to

$$
f_{PP}(y) = \pi_P \frac{\exp(-\mu) \, \mu^y}{y!}, \quad y = 1, 2, \cdots,
\tag{2}
$$

where $\pi_P = (1 - \exp(-\mu))^{-1}$. We denote $Y \sim OIPP(\mu, p)$.

Similarly, for the pmf of an OIPNB we set $f_P(\cdot)$ in (1) to

$$
f_{PNB}(y) = \pi_{NB} \frac{\Gamma(y + \kappa)}{\Gamma(\kappa)\Gamma(y + 1)} t^\kappa (1 - t)^y, \quad y = 1, 2, \cdots,
\tag{3}
$$

where $t = \frac{\kappa}{\kappa + \mu}$, $\pi_{NB} = (1 - t^\kappa)^{-1}$, and μ denotes the mean of corresponding NB distribution. We denote $Y \sim OIPNB(\kappa, \mu, p)$.

Let Y follows an OIP distribution. Some basic properties of the family of OIP distributions are shown below in which any clear proof is omitted.

i. The cumulative distribution function (cdf) is

$$F_{OIP}(y) = p + (1 - p)\pi \{F(y) - F(0)\}, \tag{4}$$

where $F(\cdot)$ is the cdf of un-truncated count distribution.

ii. Let $M(r)$ be a moment generating function (mgf) which exists for all values r and $F(0)$ be a probability at zero of a un-truncated count distribution. Then the mgf is

$$M_{OIP}(r) = pe^r + (1 - p)\pi \{M(r) - F(0)\}. \tag{5}$$

iii. Provided that $M_{OIP}(r)$ exists in a neighbourhood of $r = 0$, the m-th moment of Y is given by

$$E_{OIP}(Y^m) = p + (1 - p)\pi E(Y^m), \qquad m = 1, 2, 3, \cdots, \tag{6}$$

where it is assumed that $E(Y^m) < \infty$.

iv. For all values r which the expected value exists, the probability generating function (pgf) G of Y is defined as follows

$$G_{OIP}(r) = pr + (1 - p)\pi \{G(r) - F(0)\}, \tag{7}$$

where $G(r)$ is the pgf of un-truncated count distribution.

v. Provided that $G_{OIP}(r)$ exists in a neighbourhood of $r = 1$, the m-th factorial moment is given by

$$E_{OIP}\{(Y)_m\} = \begin{cases} p + (1 - p)\pi E(Y), & m = 1 \\ (1 - p)\pi E\{(Y)_m\}, & m = 2, 3, \cdots, \end{cases} \tag{8}$$

showing that there is a linear relationship between $E_{OIP}\{(Y)_m\}$ and the m-th factorial moment of the un-truncated distribution, e.g. $E\{(Y)_m\} = \mu^m$ for poisson and $E\{(Y)_m\} = \frac{\Gamma(\kappa + m)}{\Gamma(\kappa)}\left(\frac{\mu}{\kappa}\right)^m$ for the NB distribution.

vi. Let the recursive equation for an un-truncated distribution be

$$f(y + 1) = C(\boldsymbol{\theta}, y) f(y), \quad y = 0, 1, \cdots, \tag{9}$$

where $\boldsymbol{\theta}$ denotes corresponding parameters vector and $C(\boldsymbol{\theta}, y)$ be a known function of $\boldsymbol{\theta}$ and y. Then,

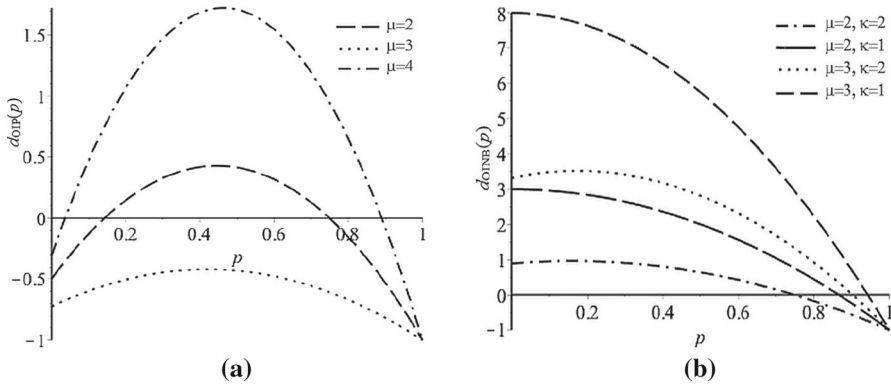$$f_P(y + 1) = C(\boldsymbol{\theta}, y) f_P(y),$$

Fig. 1 Differences between the variance and expectation of **a** the OIPP and **b** the OIPNB distributions

for $y = 1, 2, \cdots$. Similarly

$$f_{OIP}(y+1) = C(\boldsymbol{\theta}, y) f_{OIP}(y), \qquad (10)$$

for $y = 2, 3, \cdots$.

To see this, for $y = 1, 2, \cdots$, multiplying Eq. (9) by π gives a recursive equation of positive random variable. For $y = 2, 3, \cdots$, multiplying Eq. (9) by $\pi (1 - p)$ gives a recursive equation of the OIP random variable.

vii. The expectation and variance are shown to be

$$E_{OIP}(Y) = p + (1 - p)\,\pi E(Y),$$
$$Var_{OIP}(Y) = p + (1 - p)\pi \left\{ Var(Y) + E^2(Y) \right\} - E^2_{OIP}(Y), \qquad (11)$$

where $E(Y)$ and $Var(Y)$ are, respectively, the mean and variance of the underlying un-truncated distribution, e.g., $E(Y) = Var(Y) = \mu$ for poisson and $E(Y) = \mu$, $Var(Y) = \mu \left\{ 1 + \frac{\mu}{\kappa} \right\}$ for the NB distribution.

In Fig. 1 we have provided a comparison between the variance and expectation of the OIP models to illustrate evidence of how they deal with over-dispersion or under-dispersion. Figures display the difference $d_{OIP}(p) = Var_{\mathbf{OIP}}(Y) - E_{OIP}(Y)$, as a function of p, for the OIPP and OINB distributions. When the quantity $d_{OIP}(p)$ is positive (negative) then the OIP distribution carries out over-dispersion (under-dispersion). Specifically, Fig. 1a indicates that the quantity $d_{OIPP}(p)$ is positive for some values of μ over the interval $p \in (p_0, p_1]$ for some p_0 and p_1, which shows over-dispersion, whereas is negative outside of these intervals showing under-dispersion. In our case, using some numerical approaches to solve $d_{OIPP}(p) = 0$ we obtain $p_0 = 0.142, 0.036$ and $p_1 = 0.749, 0.890$ for $\mu = 3, 4$, respectively. Also, Fig. 1a illustrates that the OIPP may always handle under-dispersion for some values of μ, e.g., 2. We also observe in Fig. 1a that as p tends to 0 the OIPP becomes equivalent to the PP model which is known to display under-dispersion (Winkelmann (2008). chap. 5). Besides, Fig. 1b confirms that the

OIPNB cannot always cover over-dispersion while the proportion of ones, p, gets some large values.

Example 1 Suppose that Y follows an OIPP distribution then we obtain

$$G_{OIPP}(r) = pr + (1-p) \frac{\exp(\mu r) - 1}{\exp(\mu) - 1}, \tag{12}$$

and

$$f_{OIPP}(y+1) = \frac{\mu}{y+1} f_{OIPP}(y), \quad y = 2, 3, \cdots. \tag{13}$$

Example 2 Let Y be distributed as an OIPNB then

$$G_{OIPNB}(r) = pr + (1-p) \pi_{NB} t^\kappa \left\{ \frac{1}{[1 - r(1-t)]^\kappa} - 1 \right\}, \tag{14}$$

for $|r| < \frac{1}{1-t}$. Also,

$$f_{OIPNB}(y+1) = \frac{(\kappa + y)\mu}{(\mu + \kappa)(y+1)} f_{OIPNB}(y), \quad y = 2, 3, \cdots. \tag{15}$$

Proposition 1 *The OIPNB reduces to the OIPP distribution as $\kappa \to \infty$.*

Proof Tending κ to infinity, we obtain $\pi_{NB} \to \pi_P$, where π_{NB} and π_P are previously defined. Also, the pmf of $NB\left(\kappa, \frac{\kappa}{\kappa+\mu}\right)$ tends to $f_{pois}(\mu)$. Then Proposition (1) is proved. □

Thus the OIPP is a special case of the OIPNB distribution. Figure 2a confirms the result for $p = 0.5$ and $\mu = 1$. Using the graphical techniques we provide more comparisons between two distributions. Figure 2b shows the difference $d(y) = f_{OIPNB}(y) - f_{OIPP}(y)$, for $\kappa = 0.5, 1, 2$, $\mu = 1$ and $p = 0.5$. The comparison is performed for same means, i.e., a re-parametrization is done accordingly to make means equal. We see that the sign pattern of d(y) is $\{+, -, +\}$ as y increases on its support. The inequality $f_{OIPNB}(1) > f_{OIPP}(1)$ holds showing that for a large number of ones the OIPNB density is more appropriate than the OIPP density with the same mean. Similar pattern was additionally seen for other different values of parameters. An expectation was illustrated for κ closed to zero showing that the third sign was closer to zero.

The influence of other parameters on the probability of an one outcome is illustrated as follows. Let Y be distributed as OIPP or OIPNB, then an increase in μ reduces the probability of an one outcome and an increase in p increases the probability of an one outcome. These results were obtained readily by the differentiation of $f_{OIPNB}(1)$ (or $f_{OIPP}(1)$) with respect to all model parameters. Figure 2c, d confirm these results. By setting $\mu = \kappa = 1$, Fig. 2c shows that as p increases the probability of an one outcome increases. Similarly by putting $p = 0.5$ and $\kappa = 1$, Fig. 2d shows that as μ increases the probability of an one outcome decreases.
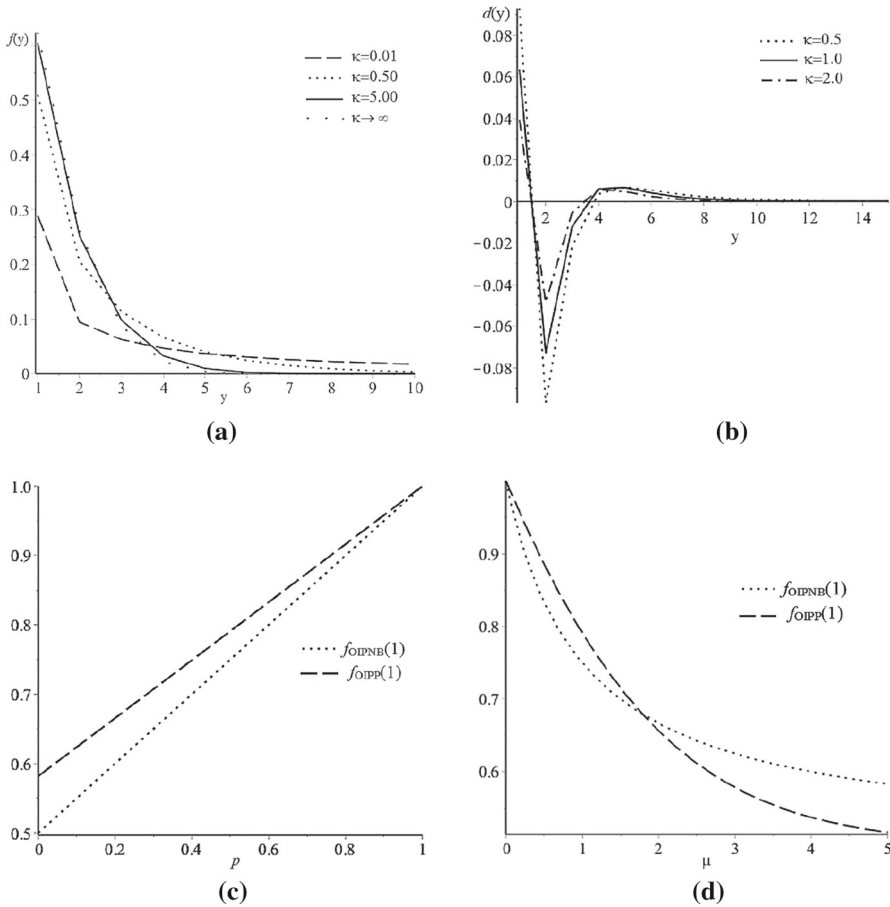
Fig. 2 **a** Probability mass functions of the OIPNB and OIPP **b** Differences between the OIPNB and OIPP densities, $d(y)$ **c** Probability of an one outcome as p increases **d** Probability of an one outcome as μ increases

3 The specification of OIP regression models

Let the Y_i $(i = 1, 2, \cdots, n)$ be independent responses, distributed as OIP and the i-th subject is associated with a m-dimensional vector $\mathbf{x}_i$ of covariates. In fitting related regression models, covariates are typically incorporated using the link function $log\,(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an m-dimensional vector of coefficients. Also a logistic scheme is used to account for predicting p_i, i.e. $log\left(\frac{p_i}{1-p_i}\right) = \mathbf{w}_i'\boldsymbol{\gamma}$, where $\mathbf{w}_i$ and $\boldsymbol{\gamma}$ denote k-dimensional vectors of covariates and coefficients, respectively. Note that the logistic models answer the question that how a covariate which induces change in μ_i affects the probability of one. The complete data likelihood is given by

$$L\,(\boldsymbol{\theta}, \boldsymbol{\gamma}; y) = \prod_{y_i=1} \{p_i + (1 - p_i)\,f_P\,(y_i)\} \prod_{y_i>1} (1 - p_i)\,f_P\,(y_i), \qquad (16)$$

where $\boldsymbol{\theta}$ denotes a parameter vector of the associated positive distribution which includes $\boldsymbol{\beta}$. Under independent sampling scheme, taking the first derivative of the log-likelihood function gives the likelihood equations

$$\frac{\partial l\left(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{y}\right)}{\partial \boldsymbol{\gamma}} = \sum_{i=1}^{n} \frac{1 - f_P\left(1\right)}{p_i + \left(1 - p_i\right) f_P\left(1\right)} p_i\left(1 - p_i\right) \mathbf{w}_i - \sum_{i=1, y_i > 1}^{n} p_i \mathbf{w}_i, \quad (17)$$

and

$$\frac{\partial l\left(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{y}\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{S_P\left(\boldsymbol{\theta}, 1\right)}{p_i + \left(1 - p_i\right) f_P\left(1\right)} + \sum_{i=1, y_i > 1}^{n} \frac{S_P\left(\boldsymbol{\theta}, y_i\right)}{f_P\left(y_i\right)}, \quad (18)$$

where $S_P\left(\boldsymbol{\theta}, \cdot\right)$ denotes the first partial derivative of the pmf with respect to $\boldsymbol{\theta}$, and P denotes any positive distribution, such as poisson and the NB. The above likelihood equations are given generally for every underlying distribution of the OIP count outcomes. Deriving the maximum likelihood estimates (MLEs) of model parameters from Eqs. (17) and (18) requires implementing advanced numerical techniques, such as Newton-Raphson. To obtain local maxima, the Hessian matrix, i.e., the matrix of second derivatives, has to be negative definite. After some algebraic operations are performed, the second derivatives of the log-likelihood, $\frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$, $\frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\gamma}}$ and $\frac{\partial^2 l(\boldsymbol{\theta}, \boldsymbol{\gamma}; \boldsymbol{y})}{\partial \boldsymbol{\gamma}^2}$ can readily be obtained but they are not tractable and available in closed forms. Thus details are omitted to save space. To solve likelihood equations simultaneously we recommend the use of optimization procedures in statistical software packages, such as NLP or NLIN in SAS, or, NLM in R. They are capable to report parameter estimates, their standard errors, and further statistical measures.

Corollary 1 *Under independent sampling scheme and setting the PP density to f_P in (16), the MLE of $\boldsymbol{\beta}$ is found by solving Eq. (18) and putting*

$$S_{Pois}\left(\boldsymbol{\beta}, y_i\right) = \left(y_i - \mu_i \pi_{i, P}\right) \mathbf{x}_i.$$

Corollary 2 *Denote P in (16) the PNB distribution and let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \kappa\right)$. Under independent sampling scheme likelihood equations are derived by setting*

$$S_{NB}\left(\kappa, y_i\right) = \frac{\left(1 - t_i\right)^2}{\mu_i} \left\{ \sum_{j=0}^{y_i - 1} \left(j + \kappa\right)^{-1} + \frac{\kappa}{t_i} - \frac{y_i}{1 - t_i} + \pi_{i, NB} \kappa t_i^{\kappa - 1} \right\}, \quad (19)$$

for parameter κ and

$$S_{NB}\left(\boldsymbol{\beta}, y_i\right) = \left\{ y_i t_i - \pi_{i, NB} \kappa \left(1 - t_i\right) \right\} \mathbf{x}_i, \quad (20)$$

for parameter $\boldsymbol{\beta}$.

Letting a logistic regression model for predicting p_i the use of advanced numerical maximization methods are required to solve the first order conditions for β, γ and κ.

Proposition 2 *Let the Y_i ($i = 1, 2, \cdots, n$) be independent count variables that each follows the OIPP or OIPNB distribution. Then an increase in the covariate $\mathbf{x}_i$ reduces the probability of an one outcome provided that the associated regression coefficient, $\boldsymbol{\beta}$, is positive, and increases it otherwise. Also, the covariate $\mathbf{z}_i$ in the logit model is directly related to the probability of an one outcome. That is, as the value of $\mathbf{z}_i$ increases (decreases) the probability of an one outcome increases (decreases) for positive $\boldsymbol{\gamma}$.*

Proof In the OIPP model

$$\frac{\partial}{\partial \mu_i} f_{OIPP}(1) = (1 - p_i)\, \pi_{i,P}^2\, exp(-\mu_i)\, \{1 - \mu_i - \exp(-\mu_i)\}, \qquad (21)$$

and in the OIPNB model

$$\frac{\partial}{\partial \mu_i} f_{OIPNB}(1) = (1 - p_i)\, t_i^\kappa (1 - t_i)\pi_{i,NB}^2 \left\{ t_i\, (\kappa + 1) - t_i^{\kappa+1} - \kappa \right\}. \qquad (22)$$

To show the effect of a change in covariate $\mathbf{x}_i$ when $\mu_i = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$, we use the chain rule which obliged the above expressions to be multiplied by $\frac{\partial}{\partial x_{ij}}\mu_i = \mu_i\beta_j$, where β_j is an element in the vector $\boldsymbol{\beta}$ corresponding to $\mathbf{x}_i$. Equations (21) and (22) are negative thus the effect at the probability of one is negative. An increase in x_{ij} reduces the probability of an one outcome if $\beta_j > 0$, and increases it otherwise. Similarly, the derivative with respect to the covariate z_{ij} for the OIPP regression model is given by

$$\frac{\partial}{\partial z_{ij}} f_{OIPP}(1) = \left\{ 1 - \pi_{i,P}\, \mu_i \exp\left(-\mu_i\right) \right\} \frac{\exp\left(-z_i'\boldsymbol{\gamma}\right)}{\left\{1 + \exp\left(-z_i'\boldsymbol{\gamma}\right)\right\}^2}\gamma_j, \qquad (23)$$

and for the OIPNB model is

$$\frac{\partial}{\partial z_{ij}} f_{OIPNB}(1) = \left\{ 1 - \pi_{i,NB}\kappa t_i^\kappa (1 - t_i) \right\} \frac{\exp\left(-z_i'\boldsymbol{\gamma}\right)}{\left\{1 + \exp\left(-z_i'\boldsymbol{\gamma}\right)\right\}^2}\gamma_j, \qquad (24)$$

where γ_j is an element in $\boldsymbol{\gamma}$. Since signs of expressions (23) and (24) are positive thus γ_j has a positive effect on the probability of one. That is, an increase in z_{ij} increases the probability of an one outcome when $\gamma_j > 0$, and reduces it otherwise. □

In below we introduce new stochastic hierarchical representations for both OIPP and OIPNB models. These are shown to be mixtures of known distributions that assist researchers to generate variants of these models or enable them to use in Bayesian approaches.

Theorem 1 *Let the Y_i, $i = 1, ..., n$, be independent counts. Consider the hierarchical representation*

$$Y_i|\omega_i, u_i \overset{ind}{\sim} poisson\left(\mu_i\left(1-\omega_i\right)+\log\left(1-u_i\mu_i^*\right)\right), \tag{25}$$

$$\omega_i \overset{iid}{\sim} Bern\left(p\right), \tag{26}$$

$$u_i \overset{iid}{\sim} Unif\left(0,1\right), \tag{27}$$

where u_i and ω_i are mutually independent and $\mu_i^* = 1 - \exp\{-\mu_i\left(1-\omega_i\right)\}$. Then the shifted count variable $Y_i + 1$ is distributed as $OIPP\left(\mu_i, p\right)$.

Proof By integrating out the u_i from the joint density of (Y_i, ω_i, u_i) we can directly show that

$$f\left(y_i|\omega_i\right) = \frac{\{\mu_i\left(1-\omega_i\right)\}^{y_i+1} e^{-\mu_i(1-\omega_i)}}{(y_i+1)!\mu_i^*}, \quad y_i = 0, 1, \cdots. \tag{28}$$

Furthermore,

$$f\left(y_i\right) = pf\left(y_i|\omega_i = 1\right) + \left(1-p\right)f\left(y_i|\omega_i = 0\right).$$

It is easy to show that $f\left(y_i|\omega_i = 1\right)$ takes 1 for $y_i = 0$ and zero otherwise and $f\left(y_i|\omega_i = 0\right)$ is of the form (28) with $\mu_i^* = 1 - \exp\left(-\mu_i\right)$. Then the proof is completed after simple statistical calculations are made. □

Theorem 2 *Suppose that the Y_i, $i = 1, ..., n$, are independent counts. Consider the hierarchical stochastic representation*

$$Y_i|v_i, \omega_i, u_i, \lambda_i \overset{ind}{\sim} Pois\left(v_i + \log\left(1-u_iv_i^*\right)\right),$$

$$v_i|\lambda_i, \omega_i \overset{ind}{\sim} Gamma\left(\kappa+1, \lambda_i+\xi_i\right), \tag{29}$$

$$\lambda_i|\omega_i \overset{ind}{\sim} F\left(\lambda_i|\omega_i\right),$$

$$\omega_i \overset{ind}{\sim} Bern\left(p\right),$$

$$u_i \overset{iid}{\sim} Unif\left(0,1\right),$$

where u_i and ω_i are mutually independent, $\xi_i = \frac{\kappa}{\mu_i(1-\omega_i)}$, $v_i^* = 1 - \exp\left(-v_i\right)$, and

$$F\left(\lambda_i|\omega_i\right) = \frac{1-\left(1+\frac{\lambda_i}{\xi_i}\right)^{-\kappa}}{1-\left(1+\frac{1}{\xi_i}\right)^{-\kappa}}, \quad 0 < \lambda_i < 1,$$

is the cdf of λ_i given ω_i. Then the shifted count variable $Y_i + 1$ is distributed as $OIPNB\left(\kappa, \mu_i, p\right)$.

Proof The marginal pmf for each Y_i, for $i = 1, \cdots, n$, is obtained directly from the joint density of variables $(Y_i, v_i, \lambda_i, u_i, \omega_i)$ by integrating out v_i, λ_i, u_i, and summing over the binary variable ω_i. First, the integration over u_i gives the explicit solution

$$f\left(y_i|v_i, \omega_i, \lambda_i\right) = \frac{v_i^{y_i+1} e^{-v_i}}{(y_i+1)! v_i^*}, \quad y_i = 0, 1, \cdots.$$

Then the conditional pmf of Y_i given ω_i is derived by integrating with respect to the distribution functions of v_i and λ_i as

$$
\begin{aligned}
f\left(y_i|\omega_i\right) &= \int_0^1 \int_0^\infty f\left(y_i|v_i, \omega_i, \lambda_i\right) dF\left(v_i|\omega_i, \lambda_i\right) dF\left(\lambda_i|\omega_i\right) \\
&= \frac{\Gamma\left(y_i+\kappa+1\right)}{\Gamma\left(y_i+2\right)\Gamma\left(\kappa\right)} \frac{1}{1-t_i^\kappa} t_i^\kappa \left(1-t_i\right)^{y_i+1},
\end{aligned}
\tag{30}
$$

for $y_i = 0, 1, \cdots$, where $t_i = \frac{\kappa}{\kappa+\mu_i(1-\omega_i)}$. Finally, we have

$$f\left(y_i\right) = p f\left(y_i|\omega_i = 1\right) + (1-p) f\left(y_i|\omega_i = 0\right).$$

It is straightforward to show that $f\left(y_i|\omega_i = 1\right) = 0$ for $y_i = 1, 2, \cdots$ and $f\left(0|\omega_i = 1\right) = 1$. Also, $f\left(y_i|\omega_i = 0\right)$ is of the form (30), where t_i reduces to $\frac{\kappa}{\kappa+\mu_i}$. After some simple algebraic operations are done we readily show that the shifted count variable $Y_i + 1$ follows $OIPNB\left(\kappa, \mu_i, p\right)$. □

Hierarchical representation (29) is useful to generate variants from the OIPNB model since one can simply generate variants from all known distributions in (29). In addition, one may use the inverse transform sampling to generate λ_i or apply the transformation $\lambda_i = \frac{(b_i-1)\kappa}{\mu_i(1-\omega_i)}$, where b_i follows the upper-truncated Pareto distribution with parameters κ, 1 and $1 + \frac{\mu_i(1-\omega_i)}{\kappa}$ (e.g. Aban et al. 2006).

4 A two-level OIP regression model

One main motivation to extend an OIP model is in the case when correlation exists between outcomes. Correlation due to clustering raises challenges. Clustering may be due to similarities among subjects. One inflation and lack of independence may occur simultaneously, which render the standard OIP model inadequate. To account the inherent correlation of subjects, we introduce a class of two-level OIP regression models with the heterogeneity effects.

Let the Y_{ij} ($i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, n_i$ and $\sum_{i=1}^n n_i = N$ gives the total number of subjects) be OIP distributed random variables of the j-th subject in the i-th cluster. In the regression setting, both $logit\left(p_{ij}\right)$ and $log\left(\mu_{ij}\right)$ are assumed to be linear functions of covariates. The covariates appearing in these two parts are not necessarily the same. In addition, random heterogeneity effects v_i and u_i for $i = 1, 2, \cdots, n$ are introduced into the linear predictors to account for possible correlation between subjects within the same cluster. These random components control unexplained variations in the model. The linear predictors are defined as

$$\log\left(\mu_{ij}\right) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i, \tag{31}$$

and

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \mathbf{z}'_{ij}\boldsymbol{\gamma} + v_i, \tag{32}$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are vectors of covariates, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding vectors of regression coefficients. Let $\mathbf{u} = (u_1, u_2, \cdots, u_n)'$ and $\mathbf{v} = (v_1, v_2, \cdots, v_n)'$. Usual assumptions are that the effects $\mathbf{u}$ and $\mathbf{v}$ are independent and distributed, respectively, as $N\left(\mathbf{0}, \sigma_u^2 \mathbf{I}_n\right)$ and $N\left(\mathbf{0}, \sigma_v^2 \mathbf{I}_n\right)$, where $\mathbf{I}_n$ denotes an $n \times n$ identity matrix. Although other distributions such as log-gamma can be adopted, but normally distributed random effects are the preferred choice in many applications. In order to capture the possible dependence between the two processes, we let the effects $\mathbf{b}_i = (u_i, v_i)$ be drawn from a bivariate distribution with density function $h(\cdot)$. The likelihood function is

$$L(\cdot) = \prod_{i=1}^{n} \int_{\mathbb{R}^2} \prod_{j=1}^{n_j} f_{OIP}\left(y_{ij}|\mathbf{b}_i\right) h\left(\mathbf{b}_i\right) d\mathbf{b}_i \tag{33}$$

No closed form solution of the underlying integral is available. With modern computing power, direct computation by numerical approximations are quite straightforward. In particular, an appropriate approximation technique to numerically evaluate the integral involved in the marginal likelihood is Gauss-Hermite quadrature. This is adopted as a useful tool in mixed modeling contexts (e.g. Liu and Pierce 1994) when a normal density is specified to the random-effects. There are several reliable procedures in standard statistical packages, such as SAS and R, to provide Gauss–Hermite quadrature calculations. Between them, proc nlmixed in SAS includes adaptive Gaussian quadrature by default to allow exact likelihood computation.

5 A simulation study

Now we conduct a simulation study to highlight the importance of accounting for one-inflation and to illustrate the impact of mis-specification on model fitting. The simulation is conducted as follows. The data set is sampled by one of the OIPNB or OIPP regression models. A total number of 1000 data sets with 500 subjects has been generated. The covariates X_1 and Z are generated independently from standard normal distributions and X_2 from a Bernoulli (0.5) distribution. For each subject, $i = 1, 2, \cdots, 500$, $\mu_i = exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$ was calculated by fixing $\beta_0 = 0.5$ and $\beta_2 = -\beta_1 = 1$ for all experiments. Also let $logit(p_i) = \gamma_0 + \gamma_1 Z_i$, and set $\gamma_0 = -0.5$ and $\gamma_1 = 0.5$. When the data generating process is according to the regression model OIPNB, we set four different values for κ, 0.2, 0.6, 1 and 2. We know as κ becomes large the OIPNB tends to OIPP and as p takes small values the distribution becomes close to the PNB distribution. Also, when simultaneously p takes small values and κ gets large then the OIPNB becomes close to the PP distribution.

Table 1 Bias and RMSE of model parameters under fitted models when the true model is OIPP

	OIPNB		OIPP		PNB		PP	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_0	0.023	0.070	0.022	0.069	−0.204	0.345	−0.349	0.354
β_1	0.002	0.029	0.001	0.028	−0.176	0.038	0.057	0.062
β_2	−0.007	0.063	−0.006	0.062	0.163	0.223	−0.043	0.073
γ_0	−0.030	0.131	−0.030	0.013				
γ_1	0.021	0.135	0.021	0.014				

Table 2 Bias and RMSE of model parameters under fitted models when the true model is OIPNB

Model		$\kappa = 0.2$		$\kappa = 0.6$		$\kappa = 1$		$\kappa = 2$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
OIPNB	β_0	−0.042	0.771	−0.028	0.270	−0.006	0.181	−0.002	0.128
	β_1	0.010	0.116	0.004	0.094	0.003	0.081	0.002	0.066
	β_2	0.001	0.113	−0.007	0.175	0.001	0.152	−0.001	0.124
	γ_0	0.001	0.177	−0.009	0.188	−0.003	0.174	−0.003	0.153
	γ_1	0.013	0.151	0.011	0.156	0.007	0.151	0.008	0.144
	κ	0.009	0.116	−0.010	0.201	0.025	0.276	0.097	0.477
OIPP	β_0	1.434	1.425	0.764	0.765	0.530	0.533	0.318	0.324
	β_1	0.241	0.242	0.218	0.220	0.167	0.169	0.105	0.111
	β_2	−0.230	0.233	−0.216	0.160	−0.166	0.148	−0.105	0.121
	γ_0	0.611	0.619	0.560	0.570	0.453	0.466	0.291	0.313
	γ_1	−0.128	0.163	−0.115	0.223	−0.094	0.176	−0.057	0.134
PNB	β_0	−4.471	4.923	−4.876	5.461	−4.712	5.445	−3.473	4.257
	β_1	−0.069	0.121	−0.137	0.170	−0.189	0.215	−0.241	0.263
	β_2	0.060	0.203	0.118	0.224	0.176	0.260	0.230	0.298
	κ	−0.196	0.197	−0.594	0.595	−0.991	0.992	−1.952	1.953
PP	β_0	0.749	0.750	0.121	0.130	−0.097	0.205	−0.064	0.082
	β_1	0.219	0.220	0.187	0.188	0.140	0.142	0.100	0.103
	β_2	−0.201	0.204	−0.174	0.181	−0.129	0.140	−0.085	0.102

However, in practical applications one may fit any of these models to positive one-inflated count data incorrectly and thus it is required to investigate the impact of this mis-specification. This importance is illustrated below.

For each replication, data are simulated according to the chosen regression model and all parameters are estimated using the PNB, PP and OIPP or OIPNB as alternative regression models. The average bias and the RMSE of estimates are obtained. Results of the simulation are summarized in Tables 1, 2, 3.

Estimation results are reported in Table 1 when the true model is specified as OIPP. Here, based on RMSE, the PP and PNB models perform poorly relative to OIPP and

Table 3 Information criterion values for various fitted models

True Model		OIPNB				OIPP
		$\kappa = 0.2$	$\kappa = 0.6$	$\kappa = 1$	$\kappa = 2$	
OIPNB	AIC	2146.49	1798.07	1721.06	1662.99	1470.22
	BIC	2171.78	1823.36	1746.35	1688.27	1495.51
OIPP	AIC	6104.59	2929.45	2413.73	2023.94	1468.40
	BIC	6125.66	2950.52	2434.80	2045.01	1489.47
PNB	AIC	2261.08	1862.27	1779.28	1734.66	1690.12
	BIC	2277.94	1879.13	1796.14	1751.52	1706.98
PP	AIC	9869.77	4484.54	3631.57	3034.54	2315.74
	BIC	9882.41	4497.18	3644.22	3047.19	2328.38

OIPNB models. The estimate of κ under the OIPNB is very large (>4E7) showing that the OIPNB reduces to the OIPP. Thus two OIPP and OIPNB models perform equally well and they have lower estimated biases with the RMSEs less than the alternative models. In overall, the average biases indicate that ignoring the one-inflated nature of data may result in under-estimation of the intercept and under- or over-estimation of the other parameters which may lead to non-significant findings.

Estimation results for the OIPP, PNB and PP fitted models are reported in Table 2 when the true model is specified as OIPNB. In this case, no alternative model is selected as the best fitted model by using the RMSE criterion. The RMSE indicates that the OIPP, PNB and PP models perform unsatisfactory relative to the OIPNB model. The bias and RMSE for all parameters of OIPP and PP models decrease when κ increases. These models under-estimated β_2 and over-estimated β_0 and β_1, whereas the PNB model over-estimated β_2 and under-estimated β_0 and β_1 in overall. The PNB model produces the most bias and RMSE for κ and the constant term. Also, in this model, for a large value of κ, the bias and RMSE of parameters (except intercept) become large. The Akaike information criterion, $AIC = -2\log L(\widehat{\theta}) + 2p$, and the Bayesian information criterion, $BIC = -2\log L(\widehat{\theta}) + p\log n$, where $L(\widehat{\theta})$, p, and n denote the likelihood, number of parameters, and sample size, respectively, are used to select the best fitted model. Smaller values of these criteria indicate better fit. In the comparison of all models, the AIC and BIC values are computed and shown in Table 3. It is seen that the previous results are confirmed by using these criteria. In fact, the OIPP and OIPNB models are comparable in fitting the OIPP model but not in the OIPNB model. This finding does suggest the importance of identifying the over-dispersion for the un-truncated distribution. Also findings indicate that the PNB and PP models are not suitable in comparison to the OIPNB and OIPP models. That is, ignoring one-inflated nature of data leads to an erroneous significant finding.

6 An empirical application with two modeling strategies

As already mentioned, the data are taken from the US national Medicare inpatient hospital database. It is emphasized in previous studies that the length of stay (LOS) in

hospital is a key element in the consumption of hospital resources and very important for hospital planning. Thus, the response variable is considered as the LOS in which, regardless of how little time a patient has spent in hospital, at least one day is credited. The data set consists 1495 observations on patients registered to a Health maintenance organization (Hmo), patient identifies themselves as Caucasian (White), patient died (Died) and type of admission (Type) variables. Type is a factor variable with three levels, elective admission (Type1), urgent admission (Type2), emergency admission (Type3). We set Type1 as the reference category. In below we report results of two regression modeling strategies to highlight the theoretical methodologies. The first is undertaken within the framework of cross-sectional data analysis which assumes that count responses are independent. The second strategy is organized through the two-level modeling setting to take into account the correlation between subjects.

6.1 Cross-sectional data analysis

We adopt the following log-linear link function

$$\log (\mu_i) = \beta_0 + \beta_1 \text{Type2}_i + \beta_2 \text{Type3}_i + \beta_3 \text{Hmo}_i + \beta_4 \text{Died}_i,$$

and the logit function

$$\log \left(\frac{p_i}{1 - p_i} \right) = \gamma_0 + \gamma_1 \text{Type2}_i + \gamma_2 \text{Type3}_i + \gamma_3 \text{Hmo}_i + \gamma_4 \text{Died}_i,$$

for $i = 1, 2, \cdots, 1495$. The variance of the LOS variable is 78.022 which is significantly larger than the mean 9.854. A fit of the PP model gives the variance 9.841 and the mean 9.854. Also the observed proportion of one counts is equal to 8.43 % while the PP model predict it in average 0.15 % which is notably less than actually observed in the data. The OIPP model predicts the probability of ones more than the PP model, however it does not account for the extra variation by receiving the variance 15.731 and the mean 9.856. For the PNB model, the estimate of the dispersion parameter κ is equal to 0.533 ($se = 0.0293$) with the corresponding p value < 0.001 which indicates strong evidence of over-dispersion. This suggests that employing the OIPNB model may be suitable to fit the data.

We fit the OIPNB regression model and make comparison to the OIPP, PP and PNB models. Table 4 shows the parameter estimates and their standard errors in fitting various models. Model performance is evaluated by using four criteria. The Hannan–Quinn information criterion, $HQIC = -2 \log L(\widehat{\theta}) + 2 p \log \{log(n)\}$, and the consistent Akaike information criterion, $CAIC = -2 \log L(\widehat{\theta}) + p \{log(n) + 1\}$, where $L(\widehat{\theta})$, p, and n denote the likelihood, number of parameters, and sample size, respectively, are also used to select the best fitted model. Smaller values of these criteria indicate better fit. We also report BIC and AIC values for each fitted model. Results show that the PP model performs very poor relative to the other models since it accommodates only positive data while the OIPNB is the preferred model as it fairly deal with extra variation of positive and one-inflated data.

Table 4 Estimates and criteria values of the OIPNB, OIPP, PNB and PP regression models

	OIPNB	OIPP	PNB	PP
β_0	2.249 (0.027)	2.298 (0.012)	2.224 (0.032)	2.264 (0.012)
β_1	0.212 (0.051)	0.218 (0.021)	0.268 (0.055)	0.250 (0.021)
β_2	0.716 (0.076)	0.708 (0.026)	0.767 (0.083)	0.750 (0.026)
β_3	−0.065 (0.054)	−0.074 (0.024)	−0.075 (0.058)	−0.076 (0.024)
β_4	−0.081 (0.043)	−0.088 (0.018)	−0.252 (0.045)	−0.249 (0.018)
κ	2.426 (0.135)		1.878 (0.103)	
γ_0	−8.091 (3.020)	−3.600 (0.217)		
γ_1	−1.200 (0.498)	−0.576 (0.283)		
γ_2	−1.007 (0.597)	−0.757 (0.449)		
γ_3	0.133 (0.314)	0.035 (0.264)		
γ_4	6.578 (3.022)	2.354 (0.238)		
AIC	9373	12,675	9487	13,703
BIC	9431	12,728	9520	13,730
HQIC	9394	12,695	9499	13,713
CAIC	9442	12,739	9525	13,735

Standard errors are in parentheses

In this analysis, the regression coefficients of the Hmo and the patient died are not statistically significant while the type of hospital admission is positively significant. Thus, patients admitted to the urgent and emergency have longer length of stay than those with the elective admission. The sign of estimates in these models is important since it can determine not only the influence of covariates on the length hospital stay in days but also can control whether a specific patient stays in hospital one day or more. This helps hospital managers to predict the number of beds at hospital. The sign of estimates for admitted patients to the urgent and emergency is positive in model part whereas is negative in the logit part. These reasonable findings show that these admitted patients stay in hospital longer than one day in comparison to those patients admitted elective. Also, the estimate of γ is positively significant in the logit part illustrating that the number of days for patient who died approaches to one day at hospital.

6.2 A two-level one-inflated regression model

The original data set includes 54 different insurance providers. All patients are nested into these providers. It is possible that some providers may provide facilities that tend to keep the length of hospital stay long. Thus, we may fit one-inflated models to take into account the correlation of patients with the same insurance provider. We also aim to see if extra variation is caused by correlation due to the existence of the provider heterogeneity in models. Table 5 presents results of fitting two-level regression models. In fitting OIP models we observed that the scale parameter of

Table 5 Estimates and criterion values from the two-level regression models

	OIPNB	OIPP	PNB	PP
β_0	2.254 (0.048)	2.295 (0.053)	2.225 (0.050)	2.242 (0.059)
β_1	0.204 (0.053)	0.211 (0.025)	0.269 (0.058)	0.263 (0.024)
β_2	0.305 (0.114)	0.151 (0.049)	0.354 (0.123)	0.155 (0.049)
β_3	-0.077 (0.056)	-0.074 (0.026)	-0.104 (0.059)	-0.094 (0.026)
β_4	-0.097 (0.043)	-0.122 (0.018)	-0.276 (0.043)	-0.275 (0.018)
κ	2.739 (0.159)		2.104 (0.120)	
γ_0	-8.068 (3.192)	-3.653 (0.227)		
γ_1	-1.221 (0.487)	-0.588 (0.290)		
γ_2	-1.361 (0.798)	-0.775 (0.462)		
γ_3	0.178 (0.343)	0.049 (0.266)		
γ_4	6.673 (3.194)	2.395 (0.248)		
σ_u^2	0.066 (0.023)	0.134 (0.030)	0.070 (0.024)	0.168 (0.037)
AIC	9333	11,974	9453	12,840
BIC	9397	12,032	9490	12,872
HQIC	9356	11,996	9467	12,852
CAIC	9409	12,043	9497	12,878

Standard errors are in parentheses

the provider heterogeneity effect in the logit part insignificant. Thus, only a random provider effect was arranged in the first part of the model. That is, for $i = 1, 2, ..., 1495$ and $j = 1, 2, ...,54$,

$$\log (\mu_{ij}) = \beta_0 + \beta_1 \text{Type2}_{ij} + \beta_2 \text{Type3}_{ij} + \beta_3 \text{Hmo}_{ij} + \beta_4 \text{Died}_{ij} + u_i,$$

where $u_i \sim N \left(0, \sigma_u^2\right)$ and

$$\log \left(\frac{p_{ij}}{1 - p_{ij}}\right) = \gamma_0 + \gamma_1 \text{Type2}_{ij} + \gamma_2 \text{Type3}_{ij} + \gamma_3 \text{Hmo}_{ij} + \gamma_4 \text{Died}_{ij}.$$

Results suggest that the two-level OIPNB model fits better than other mentioned models. In this model, the patient died is negatively significant showing that patients who died had less length of stay to persistent patients. The effect of covariates on the LOS and one day stay, or more, are rather similar to those obtained under the cross-sectional regression models while the standard errors for some of covariates are slightly increased. It appears that there is indeed an excess correlation of responses within insurance providers which takes into account extra variation. Finally, the estimate of σ_u^2 illustrates considerable heterogeneity in providers.

7 Concluding remarks

Traditional regression models for the analysis of count data are often unfitting in the presence of over-dispersion and zero-truncation issues. In the present paper, the posi-

tiveness and the one-inflation are addressed by utilizing some OIP regression models. These proposed models allow two separate mechanisms to the prediction of one-inflated counts by adopting a logit function and to the generation of subsequent counts by implementing discrete distributions, such as poisson or NB. A two-level OIP regression model is also introduced to take into account the correlation between subjects and to model excess one counts. These OIP models can provide insight into the source of excess ones and extra variation. They perform attractive alternative to zero-truncated models for count data since the corresponding marginal likelihoods are available in closed forms while the MLE of model parameters require the implementation of advanced numerical methods. The extension of our methodology straightforwardly can be applied further to various applications of count data that directly employ mixed poisson distributions with different mixing priors, such as the log-normal (e.g. Izsák 2008) or the inverse-Gaussian (Dean et al. 1989; Rigby et al. 2008). In these cases the associated marginal likelihoods are complex and need further work to make statistical inference. Moreover, hierarchical stochastic representations of two proposed models OIPP and OIPNB allow to utilize substituting enhanced estimation approaches, such as Gibbs sampling in a Bayesian perspective, that are computationally simpler than the marginal likelihood technique. These are topics of our future research.

References

Aban IB, Meerschaert MM, Panorska AK (2006) Parameter estimation for the truncated pareto distribution. J Am Stat Assoc 101(473):270–277

Cordeiro GM, Rodrigues J, de Castro M (2012) The exponential COM-poisson distribution. Stat Pap 53:653–664

Dean C, Lawless JF, Willmot GE (1989) A mixed poisson-inverse-Gaussian regression model. Can J Stat 17(2):171–181

Garay AM, Hashimotob EM, Ortega EMM, Lachos VH (2011) On estimation and influence diagnostics for zero-inflated negative binomial regression models. Comput Stat Data Anal 55:1303–1318

Gschlößl S, Czado C (2008) Modelling count data with overdispersion and spatial effects. Stat Pap 49:531–552

Gurmu S (1991) Tests for detecting over-dispersion in the positive poisson regression model. J Bus Econ Stat 9:215–222

Hall DB (2000) Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics 56:1030–1039

Hardin JW, Hilbe JM (2007) Generalized linear models and extensions, 2nd edn. Stata Press, College Station

Hilbe JM (2011) Negative binomial regression. Cambridge University Press, Cambridge

Hur K, Hedeker D, Henderson W, Khuri S, Daley J (2002) Modeling clustered count data with excess zeros in health care outcomes research. Health Serv Outcomes Res Methodol 3:5–20

Izsák R (2008) Maximum likelihood fitting of the poisson lognormal distribution. Environ Ecol Stat 15(2):143–156

Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34:1–14

Lee AH, Wang K, Yau KK, Somerford PJ (2003) Truncated negative binomial mixed regression modeling of ischaemic stroke hospitalizations. Stat Med 22(7):1129–1139

Lim HK, Li WK, Yu PLH (2014) Zero-inflated poisson regression mixture model. Comput Stat Data Anal 71:151–158

Liu Q, Pierce DA (1994) A note on Gauss–Hermite quadrature. Biometrika 81(3):624–629

Matthews JNS, Appleton DR (1993) An application of the truncated poisson distribution to immunogold a assay. Biometrics 49:617–621

Rigby RA, Stasinopoulos DM, Akantziliotou C (2008) A framework for modelling overdispersed count data, including the poisson-shifted generalized inverse Gaussian distribution. Comput Stat Data Anal 53:381–393

Sampford MR (1955) The truncated negative binomial distribution. Biometrika 42:58–69

Wang K, Yau KKW, Lee AH (2002) A zero-inflated poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. Comput Methods Prog Biomed 68:195–203

Winkelmann R (2008) Econometric analysis of count data, 5th edn. Springer, Heidelberg, New York

Xie T, Aickin MA (1997) Truncated poisson regression model with applications to occurrence of adenomatous polyps. Stat Med 16:1845–1857

Yau KK, Wang K, Lee AH (2003) Zero-inflated negative binomial mixed regression modeling of overdispersed count data with extra zeros. Bioml J 45(4):437–452

Zuur AF, Ieno EN, Neil J, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer, New York