Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros

Andy H. Lee, Kui Wang, Jane A. Scott, Kelvin K.W. Yau and Geoffrey J. McLachlan

Stat Methods Med Res 2006 15: 47
DOI: 10.1191/0962280206sm429oa

Statistical Methods in Medical Research 2006; **15**: 47–61

Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros

Andy H. Lee, Kui Wang Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, Perth, WA, Australia**, Jane A. Scott** Division of Developmental Medicine, University of Glasgow, UK**, Kelvin K.W. Yau** Department of Management Sciences, City University of Hong Kong, Hong Kong and **Geoffrey J. McLachlan** Department of Mathematics, University of Queensland, Brisbane, Qld., Australia

Count data with excess zeros relative to a Poisson distribution are common in many biomedical applications. A popular approach to the analysis of such data is to use a zero-inflated Poisson (ZIP) regression model. Often, because of the hierarchical study design or the data collection procedure, zero-inflation and lack of independence may occur simultaneously, which render the standard ZIP model inadequate. To account for the preponderance of zero counts and the inherent correlation of observations, a class of multi-level ZIP regression model with random effects is presented. Model fitting is facilitated using an expectation-maximization algorithm, whereas variance components are estimated via residual maximum likelihood estimating equations. A score test for zero-inflation is also presented. The multi-level ZIP model is then generalized to cope with a more complex correlation structure. Application to the analysis of correlated count data from a longitudinal infant feeding study illustrates the usefulness of the approach.

1 Introduction

Data with many zeros are often encountered in medical and public health studies. Failure to account for the extra zeros may result in biased parameter estimates and misleading inferences. For semi-continuous data with clumping at zero, two-part models mixing a discrete point mass (with all mass at zero) and a continuous random variable are applicable.[1] In addition to cross-sectional data, where the unit of observation is measured once, zero-inflation may also occur with repeated measures or longitudinal semi-continuous data. Mixed-distribution model with correlated random effects has been proposed to account for the simultaneous excess zeros and the correlation among measurements on the same unit of observation.[2–5]

When the non-zero part is a discrete random variable, a popular approach to analyse count data with excess zeros is to use a zero-inflated Poisson (ZIP) regression model.[6] The

Address for correspondence: Andy H. Lee Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology. GPO Box U 1987, Perth, WA, 6845, Australia. E-mail: Andy.Lee@curtin.edu.au

 10.1191/0962280206sm429oa

ZIP distribution is a mixture of the Poisson distribution and a degenerate component of point mass at zero. Its regression setting allows for covariates in both the Poisson and binary parts of the model. Böhning[7] reviewed the related literature and provided a variety of examples from different disciplines; see also Ridout *et al.*[8] for a review of the ZIP methodology. Interpretation of the ZIP analysis from a Bayesian perspective is discussed by Angers and Biswas.[9] In a medical context, a possible explanation for the excess of zeros might be due to the fact that the patient is cured after the treatment and so no realization of symptom being monitored will occur [10] p.159. Further applications of the ZIP regression model can be found in dental epidemiology,[11] occupational health,[12] child growth and development,[13] and health service research.[14] Moreover, tests for zero-inflation in count data are available in the literature.[15–17]

Often, because of the hierarchical study design or the data collection procedure, zero-inflation and lack of independence may be present simultaneously as a consequence of the inherent correlation structure and underlying heterogeneity. This is particularly prevalent in medical research where patients are typically nested within physicians or hospitals. Extensions of the ZIP model to handle clustered observations have been proposed recently. Hall,[18] Wang *et al.*[19] and Hur *et al.*[20] considered ZIP regression models with cluster-specific random effects, whereas Yau and Lee[21] incorporated distinct random effects for the Poisson and binary components of a two-part hurdle model for repeated counts. In a hurdle model, the logistic component is used to distinguish the zero and non-zero responses, whereas the non-zero observed counts are modelled via a truncated Poisson regression model. Such a conditional setting enables the interpretation of covariate effects through event incidence and frequency in the respective logistic and truncated Poisson components. Marginal models for clustered count data with excess zeros have also been developed as alternatives to the inclusion of random effects. The marginal approach either ignores the data dependency during estimation and then applies a robust sandwich estimate of the parameter variance–covariance matrix,[22] or utilizes generalized estimating equations with a dependence working correlation matrix into the model fitting algorithm.[23]

The focus of this paper is on modelling discrete hierarchical data with a preponderance of zero counts. After briefly reviewing the standard ZIP model, a multi-level ZIP regression model incorporating random effects to account for the data dependency is presented in Section 2. Model fitting procedure via an expectation-maximization (EM) algorithm is described in Section 3. In the presence of excess zeros, the technique provides better predictions than those obtained by fitting the corresponding multi-level Poisson regression model. An application in Section 4 concerning the formula feeding of infants, where the nested data collected from a longitudinal study exhibit repeated high frequency of zero counts, demonstrates the practical usefulness of the technique. Further generalizations of the ZIP methodology, including a score test for zero-inflation and an extension of the method to handle autocorrelation for serial count data, are presented in Section 5. Finally, some discussions are given in Section 6.

2 Multi-level ZIP regression model

Suppose a discrete count response variable Y follows a ZIP distribution:

$$P(Y=0) = \phi + (1-\phi)e^{-\lambda}$$
$$P(Y=y) = (1-\phi)\frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 1, 2, \ldots$$

where $0 < \phi < 1$ so that it incorporates more zeros than those permitted under the Poisson assumption ($\phi = 0$), whereas $\phi < 0$ corresponds to the zero-deflated situation.[24] The ZIP distribution may be regarded as a mixture of a Poisson (λ) and a degenerate component placing all its mass at zero. Further properties, including a graphical representation and interpretation in terms of its (unobserved) two-point heterogeneity, can be found in Böhning *et al*.[11]

For independent counts Y_j ($j = 1, \ldots, n$), Lambert[6] proposed a ZIP regression model to examine the effects of risk factors or confounders by allowing both $\log\lambda$ and the logistic transform of ϕ to be linear functions of some covariates. Maximum likelihood estimation of the regression coefficients can be performed via an EM algorithm. The model fitting procedure has been implemented in statistical packages such as STATA.[25] Recently, the ZIP regression model has been extended to the random effects setting, whereby random components w_i and u_i are incorporated within the logistic and Poisson linear predictors to account for the dependence of observations within clusters.[19,20] Similarly, a two-part conditional model was proposed to analyse repeated measures data.[21] These random effects ZIP models are cluster-specific in the sense that the random effects w_i and u_i so introduced are specific to the ith cluster. In the following, a multi-level ZIP regression model is developed to handle correlated count data with extra zeros.

Without loss of generality, consider the three-level hierarchical situation where Y_{ijk} represents the kth observation of the jth individual within the ith cluster ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n_i$; $k = 1, 2, \ldots, n_{ij}$). Let $n = \sum_{i=1}^{m} n_i$ be the total number of individuals and $N = \sum_{i=1}^{m}\sum_{j=1}^{n_i} n_{ij}$ gives the total number of observations. The observations may be taken to be independent between clusters, but certain within-cluster and within-individual correlations are anticipated, which can be modelled explicitly through random effects attached to the linear predictors:

$$\log\left[\frac{\phi_{ijk}}{(1-\phi_{ijk})}\right] = \xi_{ijk} = a_{ijk}^{\mathrm{T}}\alpha + w_i + s_{ij}$$
$$\log(\lambda_{ijk}) = \eta_{ijk} = x_{ijk}^{\mathrm{T}}\beta + u_i + v_{ij}$$

Here, the covariates a_{ijk} and x_{ijk} appearing in the respective logistic and Poisson components are not necessarily the same, and α and β are the corresponding vectors of regression coefficients. The vectors w_i and u_i denote the cluster random effects, whereas s_{ij} and v_{ij} are the random variations at subject level. Writing $w = (w_1, \ldots, w_m)^{\mathrm{T}}$,

$u = (u_1, \ldots, u_m)^T$, $s = (s_{11}, \ldots, s_{1n_1}, s_{21}, \ldots, s_{2n_2}, \ldots, s_{m1}, \ldots, s_{mn_m})^{\mathrm{T}}$, and $v = (v_{11}, \ldots, v_{1n_1}, v_{21}, \ldots, v_{2n_2}, \ldots, v_{m1}, \ldots, v_{mn_m})^{\mathrm{T}}$, the three-level ZIP regression model can be expressed in vector form as:

$$\log\left[\frac{\phi}{(1-\phi)}\right] = \xi = A\alpha + R_w w + R_s s$$
$$\log(\lambda) = \eta = X\beta + R_u u + R_v v$$

where A, X, R_w, R_s, R_u and R_v are design matrices. For simplicity of presentation, the random effects w, s, u and v are assumed to be independent and normally distributed with mean zero and variance $\sigma_w^2, \sigma_s^2, \sigma_u^2$ and σ_v^2, respectively. Although other distributions such as log-gamma can be adopted, normally distributed random effects are the preferred choice and interpretation of parameters becomes more straightforward.[20] A complex correlation structure can also be specified for the random components to accommodate the simultaneous clustering and repeated measures design, depending on the nature of the data collected and the context of the study.

3 Model estimation

Estimation of the multi-level ZIP regression model parameters can be achieved following the restricted maximum likelihood approach within the generalized linear mixed models (GLMMs) framework.[26] Construction of the GLMM penalized likelihood simply requires the log-likelihood function treating the random component as conditionally fixed, and the logarithm of the probability density function of the random effects. By assuming a multivariate normal distribution for the random effects, the corresponding probability density function can be easily constructed, thus permitting the specification of complex correlation structure in the variance components. The advantage of the GLMM approach is that estimation of parameters avoids high-dimensional integrations but only requires second-order derivatives and an appropriate iterative numerical scheme.

In the manner of Wang *et al.*,[19] the penalized log-likelihood is given by $l = l_1 + l_2$, with l_1 being the log-likelihood function when the random effects are conditionally fixed and l_2 the log density of the random effects. If the random effects are treated as parameters, then the negative of l_2 can be viewed as a penalty function on the random effects:

$$l_1 = \sum_{y_{ijk}=0} \log\left(\frac{\exp(\xi) + \exp(-\exp(\eta_{ijk}))}{1 + \exp(\xi)}\right)$$
$$+ \sum_{y_{ijk}>0} \left[y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!) - \log(1 + \exp(\xi))\right]$$

$$l_2 = -\frac{1}{2}\left[m\log\left(2\pi\sigma_u^2\right) + \sigma_u^{-2}u^{\mathrm{T}}u + n\log\left(2\pi\sigma_v^2\right) + \sigma_v^{-2}v^{\mathrm{T}}v\right]$$
$$-\frac{1}{2}\left[m\log\left(2\pi\sigma_w^2\right) + \sigma_w^{-2}w^{\mathrm{T}}w + n\log\left(2\pi\sigma_s^2\right) + \sigma_s^{-2}s^{\mathrm{T}}s\right]$$

The α coefficients can be interpreted in terms of the proportion of excess zeros, whereas the β coefficients relate to the mean response in the Poisson part. Heterogeneity among clusters and between individuals is allowed through the random effects w, u, s and v. Estimation proceeds by maximizing l_1 with the variance components fixed at their current values and then updating the values of the variance components using restricted maximum likelihood (REML) estimates obtained by consideration of l_2.[26]

To ensure the convergence and stability in the estimation of the parameters and random effects in l_1, the EM algorithm is used as adopted in Meng.[27] The EM algorithm was also used for overdispersed count data.[28] The complete-data log-likelihood is constructed as $l_C = l_\xi + l_\eta$, where

$$l_\xi = \sum_{ijk}\left(z_{ijk}\xi_{ijk} - \log(1+\exp(\xi_{ijk}))\right)$$
$$-\frac{1}{2}\left[m\log\left(2\pi\sigma_w^2\right) + \sigma_w^{-2}w^{\mathrm{T}}w + n\log\left(2\pi\sigma_s^2\right) + \sigma_s^{-2}s^{\mathrm{T}}s\right]$$
$$l_\eta = \sum_{ijk}(1-z_{ijk})\left(y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!)\right) - \frac{1}{2}[n\log(2\pi\sigma_v^2) + \sigma_v^{-2}v^{\mathrm{T}}v]$$
$$-\frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^{\mathrm{T}}u]$$

and z_{ijk} is an unobserved binary variable indicating whether y_{ijk} comes from the latent class zero ($z_{ijk}=1$) or non-zero ($z_{ijk}=0$). Treating the realization of the occurrence of extra zeros as a missing latent variable permits the decomposition of the complete-data log-likelihood l_C into two orthogonal components l_ξ and l_η, so that parameter estimation can be performed by maximizing these two functions separately. The EM algorithm proceeds by alternating between (a) replacing z_{ijk} by its conditional expectation $z_{ijk}^{(g)}$, where g denotes the gth iteration, under the current estimates $\hat{\alpha}^{(g)}$, $\hat{w}^{(g)}$, $\hat{s}^{(g)}$, $\hat{\beta}^{(g)}$, $\hat{u}^{(g)}$ and $\hat{v}^{(g)}$ (E-step):

$$z_{ijk}^{(g)} = \begin{cases} \frac{1}{1+\exp[-a_{ijk}^{\mathrm{T}}\hat{\alpha}^{(g)} - \hat{w}_i^{(g)} - \hat{s}_{ij}^{(g)} - \exp(x_{ijk}^{\mathrm{T}}\hat{\beta}^{(g)} + \hat{u}_i^{(g)} + \hat{v}_{ij}^{(g)})]} & \text{if } y_{ijk}=0 \\ 0 & \text{if } y_{ijk}\geq 1 \end{cases}$$

and then (b) with the z_{ijk}'s fixed at $z_{ijk}^{(g)}$, maximizing l_ξ and l_η (M-step) separately for $\{\hat{\alpha}^{(g+1)}, \hat{w}^{(g+1)}, \hat{s}^{(g+1)}\}$ and $\{\hat{\beta}^{(g+1)}, \hat{u}^{(g+1)}, \hat{v}^{(g+1)}\}$, in view of the orthogonal partition $l_C = l_\xi + l_\eta$. Details are given in the Appendix. The estimation procedure has been implemented as a macro in S-Plus.

4 Application

International health authorities recommend that infants be exclusively breastfed for 6 months, then introduction of complementary foods and continued breastfeeding until 12 months of age and then thereafter as long as mutually desired.[29] The introduction of supplementary bottles of formula is likely to lead to reduced breast milk production which may help explain why only 50% of women in Australia are still breastfeeding at 6 months postpartum.[30] It is thus important to determine factors affecting the frequency of formula feeds by breastfeeding women. A longitudinal infant feeding study was conducted in Perth, Australia, between September 1992 and April 1993.[31,32] At the baseline survey, information was collected on maternal age (years), parity (1=multiparous, 0=primiparous) and infant sex (1=male, 0=female), along with ethnic origin (1= Australia born, 0=elsewhere) and suburb of residence. A total of 466 breastfeeding mothers were initially recruited. The cohort was followed-up by telephone interview at 14 and 24 weeks postpartum. After excluding those stopped breastfeeding before 24 weeks and lost to follow-up, complete data were available for $n = 209$ subjects residing in $m = 15$ Perth suburbs, on which this analysis was based. The outcome variable of interest was Y=number of bottle feeds that an infant had received in the 24 h prior to the interview.

In this longitudinal study, two repeated observations were taken per individual (mother–infant pair) who were in turn nested within clusters (random suburbs in metropolitan Perth). The observed number of bottle feeds, given in Table 1, ranged from 0 to 4 (overall mean=0.11) at 14 weeks and 0 to 8 (overall mean=0.349) at 24 weeks. Assuming a separate Poisson distribution at each time point, the expected number of zeros is 187 and 147, respectively. Therefore, 6 and 33 extra infants were exclusively breastfed relative to those expected under the Poisson assumption. The zero-inflation is also evident according to the χ^2_1 score test statistics[15] of 31.7 and 149 at first and second follow-ups, respectively.

For this data set, 42% of the infants were male. The average age of women was 29 years (SD=5.2), the majority of them being multiparous (73.2%) and born in Australia (73.7%). Table 2 presents the results of fitting multi-level Poisson and multi-level ZIP

Table 1 Observed and predicted number of bottle feeds at 14 and 24 weeks from multi-level Poisson and multi-level ZIP regression models

Number of bottle feeds	Observed frequency		Predicted frequency from multi-level Poisson		Predicted frequency from multi-level ZIP	
	14 weeks	24 weeks	14 weeks	24 weeks	14 weeks	24 weeks
0	193	180	190	164	193	180
1	12	9	16	31	10	13
2	2	9	2	7	4	8
3	1	6	1	3	1	4
4	1	2	0	2	1	2
5	0	1	0	1	0	1
6	0	0	0	1	0	1
7	0	1	0	0	0	0
8	0	1	0	0	0	0

Table 2 Parameter estimates and standard errors for multi-level Poisson, multi-level ZIP and multi-level ZIP (exchangeable correlation) regression models ($m = 15$ clusters, $n = 209$ subjects, $N = 418$ observations)

	Multi-level Poisson	Multi-level ZIP		Multi-level ZIP (exchangeable correlation)	
		Logistic part	Poisson part	Logistic part	Poisson part
Intercept	−3.377 (0.977)	3.724 (1.210)	1.142 (0.815)	3.773 (1.219)	1.137 (0.774)
Maternal age	0.046 (0.033)	−0.084 (0.041)*	−0.043 (0.027)	−0.082 (0.042)*	−0.042 (0.026)
Infant sex	0.330 (0.332)	0.104 (0.411)	0.696 (0.288)*	−0.050 (0.403)	0.576 (0.268)*
Parity	−1.014 (0.367)*	0.131 (0.446)	−1.020 (0.315)*	0.329 (0.444)	−0.793 (0.290)*
Ethnic origin	−0.311 (0.356)	0.216 (0.436)	−0.015 (0.296)	0.285 (0.438)	0.050 (0.280)
Time	1.155 (0.239)*	−0.146 (0.449)	0.867 (0.340)*	−0.157 (0.445)	0.929 (0.343)*
σ^2 (cluster)	0.154	0.041	0.032	0.215	0.002
σ^2 (subject)	1.670	0.163	0.153	0.340	0.090
ρ (subject)	–	–	–	0.591	0.369

*P-value <0.05.

regression models to the hierarchical data. Under the Poisson model, only the effects of parity and time (0 = 14 weeks, 1 = 24 weeks) on bottle feeding are statistically significant after adjusting for the individual and random clustering effects. However, under the ZIP model, the infant sex also exerts significant impact on the number of bottles that an infant received at 14 weeks of age or 24 weeks of age, whereas maternal age is negatively associated with the probability of excess zeros in the logistic part.

Primiparous women gave more bottles of formula than multiparous women. This finding probably reflects a woman's confidence in her ability to provide sufficient breast milk to meet the needs of her growing infant. It is likely that primiparous women, with no prior experience of breastfeeding, will be less confident in their breast milk supply and thus tend to provide a greater number of supplementary formula feeds than more experienced mothers. Similarly, infants received a great number of bottles at 24 weeks than at 14 weeks. The significant time effect is not surprising and may be related to a mother's return to paid employment. By 24 weeks, more mothers have either returned to work or are preparing to return to work. This often necessitates the replacement of breastfeeds with bottle feeds in preparation for infants being left in the care of others.

According to the ZIP model, male infants received a higher frequency of formula feeds than female infants. This finding is consistent with the literature[33] and confirms the general perception by mothers that male infants have higher nutritional needs and should therefore receive non-breast milk fluids and foods earlier and in greater quantities than their female counterparts. The ZIP model provides additional insights on the practice of supplementary feeding, in that older mothers appear to be less likely to exclusively breastfeed, probably due to work or domestic commitments which divert their full attention from their infants.

Compared with the multi-level Poisson model, variations in the random components of the multi-level ZIP model have been substantially reduced after incorporating the logistic part to model the zero-inflation probability. It should be remarked, however, that use of the Wald or likelihood ratio test for assessing the variance parameter terms are not recommended.[20] Table 1 shows the predicted number of bottle feeds for both models, which are computed by summing the predicted probability of Y under each

model. The results suggest that the multi-level ZIP model can predict the frequency of bottle feeding better than the multi-level Poisson model, especially the number of zeros.

5 Zero-inflation test and model extensions

5.1 Score test for zero-inflation

In many applications where a preponderance of zero counts is observed, it is important to assess whether the ZIP model assumption is indeed appropriate. In the literature, tests for zero-inflation have been developed for independent data.[15–17,34] A score test for correlated count data is outlined below. The advantage of the score statistic lies in its computational convenience; only a fit of the null Poisson mixed model is required. For simplicity of presentation, we assume the random effects arise from the clustering of observations. The penalized log-likelihood function $l = l_1 + l_2$ becomes

$$l_1 = \sum_{y_{ijk}} \left[y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!) \right], \qquad l_2 = -\frac{1}{2}\left[m \log\left(2\pi\sigma_u^2\right) + \sigma_u^{-2} u^{\mathrm{T}} u \right]$$

Letting $\tau = \sigma_u^2$, the score function $U(\beta, u, \tau)$ and the Fisher information matrix $\Im(\beta, u, \tau)$ are obtained from the first and second derivatives of l with respect to β, u and τ. The score statistic for testing zero-inflation is given by

$$S(\tilde{\beta}, \tilde{u}, \tilde{\tau}) = U^{\mathrm{T}}(\tilde{\beta}, \tilde{u}, \tilde{\tau}) \tilde{\Im}^{-1} U(\tilde{\beta}, \tilde{u}, \tilde{\tau})$$

where $\tilde{\Im}$ is evaluated at the REML estimate $(\tilde{\beta}, \tilde{u}, \tilde{\tau})$. It can be shown that S has an asymptotic χ_1^2 distribution under the null hypothesis H_0: $\phi = 0$ against H_1: $\phi \neq 0$.

Simulation studies further confirm that the χ_1^2 approximation is satisfactory under a wide range of conditions.[35]

Applying the score test to the infant feeding study data, the score statistics are 9.88 at 14 weeks and 38.45 at 24 weeks postpartum, which are highly significant with reference to the asymptotic χ_1^2 distribution, providing strong evidence of zero-inflation in this correlated data set.

5.2 Multi-level ZIP regression model with autocorrelation

For the present application, data were collected at two time points (14 and 24 weeks postpartum). When repeated measures are taken over time, the resulting panel data can still be modelled within the multi-level ZIP regression framework by assuming a serial dependence correlation structure such as autoregressive process for the random effects s_{ijk} and v_{ijk}. The logistic and the Poisson parts are modelled as follows:

$$\log\left[\frac{\phi_{ijk}}{(1-\phi_{ijk})}\right] = \xi_{ijk} = a_{ijk}^{\mathrm{T}}\alpha + w_i + s_{ijk}$$

$$\log(\lambda_{ijk}) = \eta_{ijk} = x_{ijk}^{\mathrm{T}}\beta + u_i + v_{ijk}$$

With a first-order autoregressive correlation structure for random effect vectors s and v of dimension N, R_s and R_v become $N \times N$ identity matrices. Let s and v be distributed as $N(0, \sigma_s^2 B_s(\rho_s))$ and $N(0, \sigma_v^2 B_v(\rho_v))$ respectively, where $B\cdot$ represents a block diagonal matrix with block size determined by the number of serial counts for each individual and $\rho\cdot$ denotes the corresponding autocorrelation parameter for a first-order autoregressive process within each block.

The estimation procedure essentially follows that of Sections 2 and 3, with corresponding changes in the model setup, EM algorithm and variance component estimation listed below.

(1) The penalty function l_2 is replaced by

$$l_2 = -\frac{1}{2}\left[m\log\left(2\pi\sigma_u^2\right) + \sigma_u^{-2}u^{\mathrm{T}}u + N\log\left(2\pi\sigma_v^2\right) + \sigma_v^{-2}v^{\mathrm{T}}B_v^{-1}v\right]$$
$$-\frac{1}{2}\left[m\log\left(2\pi\sigma_w^2\right) + \sigma_w^{-2}w^{\mathrm{T}}w + N\log\left(2\pi\sigma_s^2\right) + \sigma_s^{-2}s^{\mathrm{T}}B_s^{-1}s\right]$$

(2) Following from procedure (1), the complete-data log-likelihood $l_{\mathrm{C}} = l_\xi + l_\eta$ is modified as

$$l_\xi = \sum_{ijk}\left(z_{ijk}\xi_{ijk} - \log(1+\exp(\xi_{ijk}))\right)$$
$$-\frac{1}{2}\left[m\log\left(2\pi\sigma_w^2\right) + \sigma_w^{-2}w^{\mathrm{T}}w + N\log\left(2\pi\sigma_s^2\right) + \sigma_s^{-2}s^{\mathrm{T}}B_s^{-1}s\right]$$
$$l_\eta = \sum_{ijk}(1-z_{ijk})\left(y_{ijk}\eta_{ijk} - \exp(\eta_{ijk}) - \log(y_{ijk}!)\right) - \frac{1}{2}[N\log(2\pi\sigma_v^2) + \sigma_v^{-2}v^{\mathrm{T}}B_v^{-1}v]$$
$$-\frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u^{\mathrm{T}}u]$$

(3) In the E-step of the EM algorithm, under the current estimates $\hat{\alpha}^{(g)}$, $\hat{w}^{(g)}$, $\hat{s}^{(g)}$, $\hat{\beta}^{(g)}$, $\hat{u}^{(g)}$ and $\hat{v}^{(g)}$:

$$z_{ijk}^{(g)} = \begin{cases} \frac{1}{1+\exp[-a_{ijk}^{\mathrm{T}}\hat{\alpha}^{(g)} - \hat{w}_i^{(g)} - \hat{s}_{ijk}^{(g)} - \exp(x_{ijk}^{\mathrm{T}}\hat{\beta}^{(g)} + \hat{u}_i^{(g)} + \hat{v}_{ijk}^{(g)})]} & \text{if } y_{ijk} = 0 \\ 0 & \text{if } y_{ijk} \geq 1 \end{cases}$$

(4) In the M-step of the EM algorithm, the required changes in the first and second derivatives of l_ξ and l_η are

$$\frac{\partial l_\xi}{\partial s} = R_s^{\mathrm{T}}\frac{\partial l_\xi}{\partial \xi} - \sigma_s^{-2}B_s^{-1}s, \qquad \frac{\partial l_\eta}{\partial v} = R_v^{\mathrm{T}}\frac{\partial l_\eta}{\partial \eta} - \sigma_v^{-2}B_v^{-1}v,$$

$$\Im_{\alpha,w,s} = \begin{bmatrix} A^{\mathrm{T}} \\ R_w^{\mathrm{T}} \\ R_s^{\mathrm{T}} \end{bmatrix} \left(-\frac{\partial^2 l_\xi}{\partial \xi \partial \xi^{\mathrm{T}}} \right) \begin{bmatrix} A & R_w & R_s \end{bmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_w^{-2} I_m & 0 \\ 0 & 0 & \sigma_s^{-2} B_s^{-1} \end{pmatrix}$$

$$\Im_{\beta,u,v} = \begin{bmatrix} X^{\mathrm{T}} \\ R_u^{\mathrm{T}} \\ R_v^{\mathrm{T}} \end{bmatrix} \left(-\frac{\partial^2 l_\eta}{\partial \eta \partial \eta^{\mathrm{T}}} \right) \begin{bmatrix} X & R_u & R_v \end{bmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_u^{-2} I_m & 0 \\ 0 & 0 & \sigma_v^{-2} B_v^{-1} \end{pmatrix}$$

(5) In the variance component estimation, the information matrix is defined as

$$\Im_{\alpha,w,s,\beta,u,v} = H + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_w^{-2} I_m & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_s^{-2} B_s^{-1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_u^{-2} I_m & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_v^{-2} B_v^{-1} \end{bmatrix}$$

(6) The estimating equations for σ_s^2 and σ_v^2 now become

$$\hat{\sigma}_s^2 = \frac{[\hat{s}^{\mathrm{T}} B_s^{-1} \hat{s} + \mathrm{tr}(B_s^{-1} V_{33})]}{N}, \qquad \hat{\sigma}_v^2 = \frac{[\hat{v}^{\mathrm{T}} B_v^{-1} \hat{v} + \mathrm{tr}(B_v^{-1} V_{66})]}{N}$$

(7) The estimating equations for ρ_s and ρ_v are given by

$$\mathrm{tr}\, B_s^{-1} \frac{\partial B_s}{\partial \rho_s} = \hat{\sigma}_s^{-2} \left[\hat{s}^{\mathrm{T}} B_s^{-1} \frac{\partial B_s}{\partial \rho_s} B_s^{-1} \hat{s} + \mathrm{tr}(V_{33} B_s^{-1} \frac{\partial B_s}{\partial \rho_s} B_s^{-1}) \right]$$

$$\mathrm{tr}\, B_v^{-1} \frac{\partial B_v}{\partial \rho_v} = \hat{\sigma}_v^{-2} \left[\hat{v}^{\mathrm{T}} B_v^{-1} \frac{\partial B_v}{\partial \rho_v} B_v^{-1} \hat{v} + \mathrm{tr}(V_{66} B_v^{-1} \frac{\partial B_v}{\partial \rho_v} B_v^{-1}) \right]$$

(8) The estimating equations in procedure (7) can be reduced to cubic equations so that the estimation of the autocorrelation parameters ρ_s and ρ_v may be achieved by numerical techniques solving the cubic equations involved.[36]

For the infant feeding study with two time points, the size of each block in the block diagonal variance–covariance matrices for the random effect vectors s and v is reduced to 2. Consequently, the autocorrelation parameter becomes an exchangeable correlation between successive observations of the same individual. Results of fitting the multi-level ZIP (with exchangeable correlation) regression model are given in Table 2. The effects of covariates are rather similar to those obtained under the multi-level ZIP model.

In addition, an appreciable correlation between observations of the same individual is found (0.591 in logistic part and 0.369 in Poisson part).

5.3 Other extensions

With appropriate modification of the probability distribution function, the method of incorporating random effects in the linear predictors can be generalized to the class of zero-modified models.[24] Although additional parameters are specified under such settings, the parameter estimation procedure essentially follows the derivations in Sections 2 and 3 and proceeds by maximizing the penalized log-likelihood via an EM algorithm. Furthermore, multi-level negative binomial or zero-inflated negative binomial regression models for overdispersed count data with extra zeros can be developed in a similar manner.[37]

6 Discussion

This paper proposes a multi-level ZIP regression model to analyse hierarchical count data containing extra zeros. The method can provide insight into the source of excess zeros and the apparent heterogeneity, while accommodating the within-cluster and within-individual correlations inherent from the data structure. Application to the longitudinal infant feeding study illustrates the usefulness of the approach. In the presence of extra zeros, the multi-level ZIP regression model enables the researchers to draw sensible and valid conclusions, in terms of identifying significant factors that affect the frequency of complementary bottle feeds, as well as distinguishing women who are likely to exclusively breastfeed from those who supplement with formula feeds. The results are logical and consistent with the breastfeeding literature.

For the multi-level ZIP regression model, estimation of parameters is facilitated using an EM algorithm in conjunction with the penalized likelihood and REML estimating equations for variance components. Section 5 outlines a score test for zero-inflation to assess the ZIP assumption and the possibility of extending the model to more complex settings. For single-level random effects ZIP models, alternative numerical integration can be applied to yield a maximum marginal likelihood solution, in which integration over the random effects distribution is approximated numerically using Gauss–Hermite quadrature.[20] However, such computations become complex with additional random effects for multi-level data.

Acknowledgements

The authors would like to thank the editor and a referee for helpful comments. This research is supported by grants from the National Health and Medical Research Council, the Australian Research Council and the Research Grants Council of Hong Kong. The model fitting procedure has been implemented as a S-Plus macro available from the corresponding author.

References

1 Lachenbruch PA. Analysis of data with excess zeros. *Statistical Methods in Medical Research* 2002; **11**: 297–302.

2 Olsen MK, Shafer JL. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* 2001; **96**: 730–45.

3 Berk KN, Lachenbruch PA. Repeated measures with zeros. *Statistical Methods in Medical Research* 2002; **11**: 303–16.

4 Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research* 2002; **11**: 341–55.

5 Yau KKW, Lee AH, Ng ASK. A zero-augmented gamma mixed model to analyse longitudinal data with many zeros. *Australia and New Zealand Journal of Statistics* 2002; **44**: 177–83.

6 Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.

7 Böhning D. Zero-inflated Poisson models and C.A. Man: a tutorial collection of evidence. *Biometrical Journal* 1998; **40**: 833–43.

8 Ridout M, Demétrio CGB, Hinde J. Models for count data with many zeros. *Proceedings of the XIXth International Biometrics Conference*, Cape Town, 1998: 179–92.

9 Angers JF, Biswas A. A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics and Data Analysis* 2003; **42**: 37–46.

10 McLachlan GJ, Peel D. *Finite mixture models*. Wiley, 2000.

11 Böhning D, Dietz E, Schlattmann P, Mendonca L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society A* 1999; **162**: 195–209.

12 Lee AH, Wang K, Yau KKW. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* 2001; **43**: 963–75.

13 Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in Medicine* 2002; **21**: 1461–9.

14 Lee AH, Xiang L, Fung WK. Sensitivity of score tests for zero-inflation in count data. *Statistics in Medicine* 2004; **23**: 2757–69.

15 Van den Broek J. A score test for zero-inflation in a Poisson distribution. *Biometrics* 1995; **51**: 738–43.

16 Ridout M, Hinde J, Demétrio CGB. A score test for testing zero inflated Poisson regression model against zero inflated negative binomial alternatives. *Biometrics* 2001; **57**: 219–23.

17 Jansakul N, Hinde JP. Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis* 2002; **40**: 75–96.

18 Hall DB. Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* 2000; **56**: 1030–9.

19 Wang K, Yau KKW, Lee AH. A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. *Computer Methods and Programs in Biomedicine* 2002; **68**: 195–203.

20 Hur K, Hedeker D, Henderson W, Khuri S, Daley J. Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology* 2002; **3**: 5–20.

21 Yau KKW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine* 2001; **20**: 2907–20.

22 Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research* 2002; **11**: 317–25.

23 Dobbie MJ, Welsh AH. Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics* 2001; **43**: 431–44.

24 Dietz K, Böhning D. On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis* 2000; **34**: 441–59.

25 Stata Corporation. *Stata Statistical Software*. Release 8. Stata corporation, 2003.

26 McGilchrist CA. Estimation in generalized mixed models. *Journal of the Royal Statistical Society B* 1994; **56**: 61–9.

27 Meng XL. The EM algorithm and medical studies: A historical link. *Statistical Methods in Medical Research* 1997; **6**: 3–23.

28 McLachlan GJ. On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research* 1997; **6**: 76–98.
29 World Health Organization. *The optimal duration of breastfeeding*. Geneva: WHO, 2001.
30 Donath A, Amir S. Rates of breastfeeding in Australia by State and socio-economic status: evidence from the 1995 National Health Survey. *Journal of Paediatrics and Child Health* 2000; **36**: 164–8.
31 Scott J, Binns C, Aroni R. The influence of reported paternal attitudes on the decision to breast-feed. *Journal of Paediatrics and Child Health* 1997; **33**: 305–7.
32 Scott J, Aitkin I, Binns C, Aroni R. Factors associated with the duration of breast-feeding amongst women in Perth, Australia. *Acta Paediatrica* 1999; **88**: 416–21.
33 Pande H, Unwin C, Haheim L. Factors associated with the duration of breastfeeding: analysis of the primary and secondary responders to a self-completed questionnaire. *Acta Paediatrica* 1997; **86**: 173–7.
34 Deng D, Paul SR. Score tests for zero inflation in generalized linear models. *The Canadian Journal of Statistics* 2000; **27**: 563–70.
35 Xiang L, Lee AH, Yau KKW, McLachlan GJ. A score test for zero-inflation in correlated count data. *Statistics in Medicine* (in press).
36 Yau KKW, McGilchrist CA. ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine* 1998; **17**: 1201–13.
37 Yau KKW, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal* 2003; **45**: 437–52.

Appendix A

A.1 M-step of the EM algorithm

With the variance parameters (σ_w^2, σ_s^2) and (σ_u^2, σ_v^2) held fixed, the M-step of the EM algorithm provides estimates of the parameters and is performed using the following two sets of recursive equations:

$$\begin{bmatrix}\hat{\alpha}\\ \hat{w}\\ \hat{s}\end{bmatrix} = \begin{bmatrix}\alpha_0\\ w_0\\ s_0\end{bmatrix} + \Im_{\alpha,w,s}^{-1}\begin{bmatrix}\dfrac{\partial l_\xi}{\partial \alpha}\\ \dfrac{\partial l_\xi}{\partial w}\\ \dfrac{\partial l_\xi}{\partial s}\end{bmatrix}, \qquad \begin{bmatrix}\hat{\beta}\\ \hat{u}\\ \hat{v}\end{bmatrix} = \begin{bmatrix}\beta_0\\ u_0\\ v_0\end{bmatrix} + \Im_{\beta,u,v}^{-1}\begin{bmatrix}\dfrac{\partial l_\eta}{\partial \beta}\\ \dfrac{\partial l_\eta}{\partial u}\\ \dfrac{\partial l_\eta}{\partial v}\end{bmatrix}$$

where $\{\alpha_0, w_0, s_0\}$ and $\{\beta_0, u_0, v_0\}$ are initial values of the parameters and are replaced by their updated estimates in each iteration. The first and second derivatives of l_ξ for the logistic part are given by

$$\frac{\partial l_\xi}{\partial \alpha} = A^{\mathrm{T}}\frac{\partial l_\xi}{\partial \xi}, \qquad \frac{\partial l_\xi}{\partial w} = R_w^{\mathrm{T}}\frac{\partial l_\xi}{\partial \xi} - \sigma_w^{-2}w, \qquad \frac{\partial l_\xi}{\partial s} = R_s^{\mathrm{T}}\frac{\partial l_\xi}{\partial \xi} - \sigma_s^{-2}s$$

$$\Im_{\alpha,w,s} = \begin{bmatrix}A^{\mathrm{T}}\\ R_w^{\mathrm{T}}\\ R_s^{\mathrm{T}}\end{bmatrix}\left(-\frac{\partial^2 l_\xi}{\partial \xi\, \partial \xi^{\mathrm{T}}}\right)\begin{bmatrix}A & R_w & R_s\end{bmatrix} + \begin{pmatrix}0 & 0 & 0\\ 0 & \sigma_w^{-2}I_m & 0\\ 0 & 0 & \sigma_s^{-2}I_n\end{pmatrix}$$

where $\partial l_\xi/\partial\xi = z - \exp(\xi)/(1+\exp(\xi))$ and $\partial^2 l_\xi/\partial\xi\partial\xi^{\mathrm{T}} = \mathrm{Diag}[-\exp(\xi)/(1+\exp(\xi))^2]$.
The first and second derivatives of l_η for Poisson part are given by

$$\frac{\partial l_\eta}{\partial\beta} = X^{\mathrm{T}}\frac{\partial l_\eta}{\partial\eta}, \qquad \frac{\partial l_\eta}{\partial u} = R_u^{\mathrm{T}}\frac{\partial l_\eta}{\partial\eta} - \sigma_u^{-2}u, \qquad \frac{\partial l_\eta}{\partial v} = R_v^{\mathrm{T}}\frac{\partial l_\eta}{\partial\eta} - \sigma_v^{-2}v$$

$$\Im_{\beta,u,v} = \begin{bmatrix} X^{\mathrm{T}} \\ R_u^{\mathrm{T}} \\ R_v^{\mathrm{T}} \end{bmatrix}\left(-\frac{\partial^2 l_\eta}{\partial\eta\partial\eta^{\mathrm{T}}}\right)\begin{bmatrix} X & R_u & R_v \end{bmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma_u^{-2}I_m & 0 \\ 0 & 0 & \sigma_v^{-2}I_n \end{pmatrix}$$

where $\partial l_\eta/\partial\eta = (1-z)(y-\exp(\eta))$ and $\partial^2 l_\eta/\partial\eta\partial\eta^{\mathrm{T}} = \mathrm{Diag}[-(1-z)\ \exp(\eta)]$.

A.2 Variance component estimation

Estimation of variance of random effects requires the calculation of the information matrix. The expectation of the second derivatives are

$$\begin{aligned} E\left[\frac{\partial^2 l_1}{\partial\xi\partial\xi^{\mathrm{T}}}\right] &= E\left[\mathrm{Diag}\left(I(y=0)\frac{\exp(\xi-\exp(\eta))}{[\exp(\xi)+\exp(-\exp(\eta))]^2} - \frac{\exp(\xi)}{(1+\exp(\xi))^2}\right)\right] \\ &= \mathrm{Diag}\left(\frac{\exp(\xi-\exp(\eta))}{(1+\exp(\xi))[\exp(\xi)+\exp(-\exp(\eta))]} - \frac{\exp(\xi)}{(1+\exp(\xi))^2}\right) \end{aligned}$$

where $I(y=0)$ is set to 1 if $y=0$ and zero otherwise.

$$\begin{aligned} E\left[\frac{\partial^2 l_1}{\partial\xi\partial\eta^{\mathrm{T}}}\right] &= E\left[\mathrm{Diag}\left(I(y=0)\frac{\exp(\xi+\eta-\exp(\eta))}{[\exp(\xi)+\exp(-\exp(\eta))]^2}\right)\right] \\ &= \mathrm{Diag}\left(\frac{\exp(\xi+\eta-\exp(\eta))}{(1+\exp(\xi))[\exp(\xi)+\exp(-\exp(\eta))]}\right) \end{aligned}$$

$$\begin{aligned} E\left[\frac{\partial^2 l_1}{\partial\eta\partial\eta^{\mathrm{T}}}\right] &= -E\left[\mathrm{Diag}\left(\exp(\eta) - I(y=0)\frac{\exp(\xi+\eta)[\exp(\xi)+\exp(-\exp(\eta)) + \exp(\eta-\exp(\eta))]}{[\exp(\xi)+\exp(-\exp(\eta))]^2}\right)\right] \\ &= \mathrm{Diag}\left(\exp(\eta) - \frac{\exp(\xi+\eta)[\exp(\xi)+\exp(-\exp(\eta)) + \exp(\eta-\exp(\eta))]}{(1+\exp(\xi))[\exp(\xi)+\exp(-\exp(\eta))]}\right) \end{aligned}$$

Then the information matrix is given by

$$\Im_{\alpha,w,s,\beta,u,v} = H + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_w^{-2} I_m & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_s^{-2} I_n & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_u^{-2} I_m & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_v^{-2} I_n \end{bmatrix}$$

where $H = \begin{bmatrix} A^{\mathrm{T}} & 0 \\ R_w^{\mathrm{T}} & 0 \\ R_s^{\mathrm{T}} & 0 \\ 0 & X^{\mathrm{T}} \\ 0 & R_u^{\mathrm{T}} \\ 0 & R_v^{\mathrm{T}} \end{bmatrix} \begin{pmatrix} E\left[-\frac{\partial^2 l_1}{\partial \xi \partial \xi^{\mathrm{T}}}\right] & E\left[-\frac{\partial^2 l_1}{\partial \xi \partial \eta^{\mathrm{T}}}\right] \\ E\left[-\frac{\partial^2 l_1}{\partial \eta \partial \xi^{\mathrm{T}}}\right] & E\left[-\frac{\partial^2 l_1}{\partial \eta \partial \eta^{\mathrm{T}}}\right] \end{pmatrix} \begin{bmatrix} A & R_w & R_s & 0 & 0 & 0 \\ 0 & 0 & 0 & X & R_u & R_v \end{bmatrix}$

Denoting the inverse of $\Im_{\alpha,w,s,\beta,u,v}$ by $V = (V_{ij})$, $i = 1, \ldots, 6$ and $j = 1, \ldots, 6$, so that V_{22}, V_{33}, V_{55} and V_{66} are block matrices corresponding to random effects w, s, u and v, respectively. Then

$$\hat{\sigma}_w^2 = \frac{[\hat{w}^{\mathrm{T}} \hat{w} + \mathrm{tr}(V_{22})]}{m}, \quad \hat{\sigma}_s^2 = \frac{[\hat{s}^{\mathrm{T}} \hat{s} + \mathrm{tr}(V_{33})]}{n}$$

$$\hat{\sigma}_u^2 = \frac{[\hat{u}^{\mathrm{T}} \hat{u} + \mathrm{tr}(V_{55})]}{m}, \quad \hat{\sigma}_v^2 = \frac{[\hat{v}^{\mathrm{T}} \hat{v} + \mathrm{tr}(V_{66})]}{n}$$

Finally, the square root of diagonal elements of the block matrices V_{11} and V_{44} provides the respective standard errors for the estimates of regression coefficients α and β.