STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1608–1622
Published online 22 June 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2616

A score test for overdispersion in zero-inflated poisson mixed regression model

Liming Xiang[1], Andy H. Lee[2], Kelvin K. W. Yau[1,*,†] and Geoffrey J. McLachlan[3]

[1]*Department of Management Sciences, City University of Hong Kong, Hong Kong*
[2]*Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, Perth, Australia*
[3]*Department of Mathematics, University of Queensland, Brisbane, Australia*

SUMMARY

Count data with extra zeros are common in many medical applications. The zero-inflated Poisson (ZIP) regression model is useful to analyse such data. For hierarchical or correlated count data where the observations are either clustered or represent repeated outcomes from individual subjects, a class of ZIP mixed regression models may be appropriate. However, the ZIP parameter estimates can be severely biased if the non-zero counts are overdispersed in relation to the Poisson distribution. In this paper, a score test is proposed for testing the ZIP mixed regression model against the zero-inflated negative binomial alternative. Sampling distribution and power of the test statistic are evaluated by simulation studies. The results show that the test statistic performs satisfactorily under a wide range of conditions. The test procedure is applied to pancreas disorder length of stay that comprised mainly same-day separations and simultaneous prolonged hospitalizations. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: count data; negative binomial; overdispersion; random effects; score test; zero-inflation

1. INTRODUCTION

In many medical applications, the count data encountered contain excess zeros relative to the Poisson distribution. A popular approach to analyse such data is to use a zero-inflated Poisson (ZIP) regression model [1]. The ZIP model combines the Poisson distribution with a degenerate component of point mass at zero. A review of the ZIP methodology can be found in References

*Correspondence to: Kelvin K. W. Yau, Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong.
†E-mail: mskyau@cityu.edu.hk

Contract/grant sponsor: Australian Research Council
Contract/grant sponsor: Research Grants Council of Hong Kong
Contract/grant sponsor: Health Information Centre, Department of Health, Western Australia

Received 2 November 2005
Accepted 5 May 2006

[2, 3]. Often, zero-inflation and dependency are present simultaneously due to the hierarchical study design or longitudinal data collection procedure. The inherent correlation is common in medical research where patients are typically nested within hospitals or health regions. Extensions of the ZIP regression model have been developed, in which random effects are incorporated within the Poisson and binary components of the ZIP model to handle the clustered heterogeneous counts [4–7]. Recently, a class of multi-level ZIP regression model with random effects is proposed [8]. Model fitting is facilitated using an EM algorithm [9], while variance components are estimated via residual maximum likelihood estimating equations. Marginal models utilizing generalized estimating equations have also been suggested as alternatives to the inclusion of random effects [10, 11].

Prior to application of the standard ZIP model, it is important to assess whether the ZIP assumption is indeed valid. Score tests for zero-inflation in count data are available in the literature [12, 13], with extensions in more generalized settings [14, 15] and specific applications such as disease mapping by parametric bootstrap [16]. Sensitivity of score tests for zero-inflation have also been considered [17]. For correlated count data, a score test for zero-inflation testing the Poisson mixed model against its ZIP mixed counterpart is appropriate [18]. The advantage of the score statistic lies in its computational convenience; only a fit of the null Poisson mixed regression model is required. The test procedure has been applied to analyse recurrent urinary track infections in elderly women, where the correlated data collected from a retrospective cohort study exhibit a preponderance of zero counts [18].

In practice, count data are often overdispersed so that alternative distributions such as the zero-inflated negative binomial (ZINB) may be more appropriate than the ZIP. Moreover, it has been established that the ZIP parameter estimates can be inconsistent in the event of severe overdispersion for the non-zero counts [19]. Consequently, Ridout, Hinde and Demetrio provided a score test for testing a ZIP regression model against a ZINB alternative, based on a general parameterization of the negative binomial distribution [19]. However, no equivalent test is available for correlated count data exhibiting zero-inflation as well as overdispersion.

In this paper, a score test is presented for assessing overdispersion in the ZIP mixed regression model against a ZINB mixed alternative. The development parallels that of Ridout *et al.* [19]. After briefly reviewing the ZIP mixed and ZINB mixed regression models in Section 2, the score test for overdispersion and corresponding hypotheses are specified in Section 3. The sampling distribution of the score test statistic and its power properties are investigated via simulation experiments in Section 4. To illustrate the test procedure, an example on pancreas disorder length of stay is provided in Section 5, where same-day separations are becoming more prevalent and the heterogeneous count data collected from the same hospital are correlated. Finally, some concluding remarks are given in Section 6.

2. ZIP AND ZINB MIXED REGRESSION MODELS

Let Y denote the count variable of interest. Suppose the jth response variable from the ith cluster follows a ZIP distribution:

$$P(Y_{ij}=y_{ij})=\begin{cases}\omega_{ij}+(1-\omega_{ij})\exp(-\lambda_{ij}), & y_{ij}=0\\ (1-\omega_{ij})\exp(-\lambda_{ij})\lambda_{ij}^{y_{ij}}/y_{ij}!, & y_{ij}>0\end{cases}$$

DOI: 10.1002/sim

$i=1,\ldots,m$ and $j=1,\ldots,n_i$, where m is the number of clusters and n_i is the number of observations within cluster i. The mean and variance of the ZIP random variables are given by:

$$E(Y_{ij})=(1-\omega_{ij})\lambda_{ij}$$

$$\text{var}(Y_{ij})=(1-\omega_{ij})\lambda_{ij}(1+\omega_{ij}\lambda_{ij})$$

In the regression setting, both the mean λ_{ij} and zero proportion ω_{ij} parameters are related to the covariate vectors $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, respectively. Moreover, responses within the same cluster/subject are likely to be correlated. To accommodate the inherent correlation, random effects u_i and v_i are incorporated in the linear predictors η_{ij} for the Poisson part and ξ_{ij} for the zero mixing part. The ZIP mixed regression model is thus [6]:

$$\eta_{ij}=\log(\lambda_{ij})=\mathbf{x}'_{ij}\beta+u_i$$

$$\xi_{ij}=\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right)=\mathbf{z}'_{ij}\gamma+v_i$$

where β and γ are the corresponding $p\times 1$ and $q\times 1$ vector of regression coefficients. The random effects u_i and v_i are assumed to be independent and normally distributed with mean 0 and variance σ_u^2 and σ_v^2, respectively.

On the other hand, the ZINB model features the modelling of the observed overdispersion via the negative binomial component besides accounting for the excess zeros. The count variable Y_{ij} follows a ZINB distribution of the form:

$$P(Y_{ij}=y_{ij})=\begin{cases}\omega_{ij}+(1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}, & y_{ij}=0\\ (1-\omega_{ij})\dfrac{\Gamma(y_{ij}+(1/\alpha))}{y_{ij}!\Gamma(1/\alpha)}(1+\alpha\lambda_{ij})^{-1/\alpha}\left(1+\dfrac{1}{\alpha\lambda_{ij}}\right)^{-y_{ij}}, & y_{ij}>0\end{cases}$$

where $\alpha>0$ is a dispersion parameter. The mean and variance of Y_{ij} are:

$$E(Y_{ij})=(1-\omega_{ij})\lambda_{ij}$$

$$\text{var}(Y_{ij})=(1-\omega_{ij})\lambda_{ij}(1+\omega_{ij}\lambda_{ij}+\alpha\lambda_{ij})$$

Note that the ZINB distribution reduces to the ZIP distribution in the limit $\alpha\to 0$.

Analogous to the ZIP mixed regression model, a ZINB mixed regression model can be defined [20]. For simplicity of presentation, let $N=\sum_{i=1}^{m}n_i$, $u=(u_1,\ldots,u_m)'$, $v=(v_1,\ldots,v_m)'$ be random effects vectors; $\eta=(\eta_{11},\ldots,\eta_{1n_1},\ldots,\eta_{m1},\ldots,\eta_{mn_m})'$, $\xi=(\xi_{11},\ldots,\xi_{1n_1},\ldots,\xi_{m1},\ldots,\xi_{mn_m})'$. The $N\times 1$ vectors λ and ω are defined in a similar manner. Also, let the $N\times p$, $N\times q$ and $N\times m$ design matrices be partitioned as:

$$\mathbf{X}=[\mathbf{x}_{11},\ldots,\mathbf{x}_{1n_1},\ldots,\mathbf{x}_{m1},\ldots,\mathbf{x}_{mn_m}]'$$

$$\mathbf{Z}=[\mathbf{z}_{11},\ldots,\mathbf{z}_{1n_1},\ldots,\mathbf{z}_{m1},\ldots,\mathbf{z}_{mn_m}]'$$

$$\mathbf{W}=\text{diag}[\mathbf{1}_{n_1},\mathbf{1}_{n_2}\ldots,\mathbf{1}_{n_m}]=[\mathbf{w}_{11},\ldots,\mathbf{w}_{1n_1},\ldots,\mathbf{w}_{m1},\ldots,\mathbf{w}_{mn_m}]'$$

DOI: 10.1002/sim

where $\mathbf{1}_n$ denotes an $n \times 1$ vector of 1. Then the ZINB mixed regression model can be expressed in matrix notation as:

$$\log(\lambda) = \eta = \mathbf{X}\beta + \mathbf{W}u$$

$$\log\left(\frac{\omega}{1-\omega}\right) = \xi = \mathbf{Z}\gamma + \mathbf{W}v$$

Based on the generalized linear mixed model formulation [21], the residual maximum likelihood (REML) estimates of the ZINB mixed regression model parameters can be obtained. The approach commences with the best linear unbiased prediction-type log-likelihood $l = l_1 + l_2$, where l_1 and l_2 represent the respective contribution of the fixed and random effects:

$$l_1 = \sum_{i,j} I_{\{y_{ij}=0\}} \log(\omega_{ij} + (1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}) + I_{\{y_{ij}>0\}} \left[\log(1-\omega_{ij}) - \log(y_{ij}!) + \log\left(\Gamma\left(y_{ij}+\frac{1}{\alpha}\right)\right) - \log\left(\Gamma\left(\frac{1}{\alpha}\right)\right) - \left(y_{ij}+\frac{1}{\alpha}\right)\log(1+\alpha\lambda_{ij}) + y_{ij}\log(\alpha\lambda_{ij})\right]$$

$$l_2 = -\frac{1}{2}[m\log(2\pi\sigma_u^2) + \sigma_u^{-2}u'u + m\log(2\pi\sigma_v^2) + \sigma_v^{-2}v'v]$$

with indicator function $I_{(\cdot)}$ being 1 if the specified condition is satisfied and 0 otherwise. Here l can be viewed as a penalized log-likelihood function with l_2 being the penalty for the ZINB log-likelihood l_1 when the random effects are conditionally fixed. Let the parameter vector of interest be $\boldsymbol{\theta} = (\alpha, \beta, \gamma, u, v)'$. With suitable initial values, the REML estimates of $\boldsymbol{\theta}$ can be obtained iteratively by maximizing l, via an EM algorithm to ensure convergence. The variance component estimates for σ_u^2 and σ_v^2 are then computed from estimating equations involving $\boldsymbol{\theta}$; details of the estimation procedure can be found in Reference [20]. The ZIP mixed regression model may be considered as a special case of the ZINB mixed regression model when $\alpha \to 0$, the corresponding parameter estimates are obtained in a similar manner via an EM algorithm and associated REML estimating equations [6].

3. SCORE TEST FOR OVERDISPERSION

To test for overdispersion in the ZIP mixed regression model against the ZINB mixed regression model is equivalent to testing the null hypothesis $H_0 : \alpha = 0$ against the alternative $H_1 : \alpha > 0$. Our development of the score test parallels that of Ridout *et al.* for independent data [19]. Derivation of the score test requires the first- and second-order derivatives of l with respective to α, β, γ, u and v, and then evaluated at REML estimates of the ZIP mixed regression model, i.e. under the null hypothesis. Details of the derivatives involved are given in Appendix A.

First, the efficient score S is obtained by evaluating the derivative of l with respective to α at the REML estimates $\hat{\beta}$, $\hat{\gamma}$, $\hat{u}$ and $\hat{v}$ of the ZIP mixed regression model:

$$S = \frac{1}{2}\sum_{i,j}\left\{[(y_{ij}-\hat{\lambda}_{ij})^2 - y_{ij}] - I_{(y_{ij}=0)}\frac{\hat{\lambda}_{ij}^2\hat{\omega}_{ij}}{\hat{p}_{0,ij}}\right\}$$

DOI: 10.1002/sim

where the probability $\hat{p}_{0,ij}=\hat{\omega}_{ij}+(1-\hat{\omega}_{ij})\exp(-\hat{\lambda}_{ij})$. The derivation of this score S is in principle the same as the score statistic Z_2 proposed by Deng and Paul [22] which does not involve random effects. In the following, all expectations are taken under $H_0:\alpha=0$ and conditional on random effects u and v. Based on the second derivatives of l evaluated at the REML estimates, the expected Fisher information matrix is partitioned as

$$\mathbf{J}=E\left(-\frac{\partial^2 l}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)=\begin{bmatrix} J_{\alpha\alpha} & J_{\alpha\beta'} & J_{\alpha\gamma'} & J_{\alpha u'} & J_{\alpha v'} \\ J_{\alpha\beta} & & & & \\ J_{\alpha\gamma} & & \mathbf{J}_{(\beta,\gamma,u,v)} & & \\ J_{\alpha u} & & & & \\ J_{\alpha v} & & & & \end{bmatrix}$$

with

$$\mathbf{J}_{(\beta,\gamma,u,v)}=\begin{bmatrix} \mathbf{X}' & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}' \\ \mathbf{W}' & \mathbf{0} \\ \mathbf{0} & \mathbf{W}' \end{bmatrix}\begin{bmatrix} E\left(-\frac{\partial^2 l}{\partial\eta\partial\eta'}\right) & E\left(-\frac{\partial^2 l}{\partial\eta\partial\xi'}\right) \\ E\left(-\frac{\partial^2 l}{\partial\xi\partial\eta'}\right) & E\left(-\frac{\partial^2 l}{\partial\xi\partial\xi'}\right) \end{bmatrix}\begin{bmatrix} \mathbf{X} & \mathbf{0} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} & \mathbf{0} & \mathbf{W} \end{bmatrix}$$

$$+\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{\sigma}_{\mathbf{u}}^{-\mathbf{2}}\mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \hat{\sigma}_{v}^{-\mathbf{2}}\mathbf{I}_m \end{bmatrix}$$

where $\mathbf{I}_m$ denotes an $m\times m$ identity matrix. Here,

$$E\left(-\frac{\partial^2 l}{\partial\eta\partial\eta'}\right)=\mathrm{diag}[\hat{\lambda}\{(1-\hat{\omega})-\hat{\lambda}\hat{\omega}(1-\hat{\omega}/\hat{p}_0)\}]$$

$$E\left(-\frac{\partial^2 l}{\partial\xi\partial\xi'}\right)=\mathrm{diag}[-\hat{\omega}^2(1/\hat{p}_0-1)]$$

$$E\left(-\frac{\partial^2 l}{\partial\eta\partial\xi'}\right)=\mathrm{diag}[-\hat{\lambda}\hat{\omega}(1-\hat{\omega}/\hat{p}_0)]$$

and $\hat{p}_0$ is a vector with elements $\hat{p}_{0,ij}$. Typical entries of $\mathbf{J}$ are as follows:

$$J_{\alpha\alpha}=\frac{1}{2}\sum_{i,j}\left\{(1-\hat{\omega}_{ij})\hat{\lambda}_{ij}^2-\frac{1}{4}\hat{\omega}_{ij}\hat{\lambda}_{ij}^4\left(1-\frac{\hat{\omega}_{ij}}{\hat{p}_{o,ij}}\right)\right\}$$

DOI: 10.1002/sim

$$J_{\alpha\beta} = \frac{1}{2}\sum_{i,j}\left\{\hat{\lambda}_{ij}^3\hat{\omega}_{ij}\left(1-\frac{\hat{\omega}_{ij}}{\hat{p}_{o,ij}}\right)\mathbf{x}_{ij}\right\}$$

$$J_{\alpha\gamma} = \frac{1}{2}\sum_{i,j}\left\{\hat{\lambda}_{ij}^2\hat{\omega}_{ij}\left(1-\frac{\hat{\omega}_{ij}}{\hat{p}_{o,ij}}\right)\mathbf{z}_{ij}\right\}$$

$$J_{\alpha u} = \frac{1}{2}\sum_{i,j}\left\{\hat{\lambda}_{ij}^3\hat{\omega}_{ij}\left(1-\frac{\hat{\omega}_{ij}}{\hat{p}_{o,ij}}\right)\mathbf{w}_{ij}\right\}$$

$$J_{\alpha v} = \frac{1}{2}\sum_{i,j}\left\{\hat{\lambda}_{ij}^2\hat{\omega}_{ij}\left(1-\frac{\hat{\omega}_{ij}}{\hat{p}_{o,ij}}\right)\mathbf{w}_{ij}\right\}$$

The score statistic for testing overdispersion in the ZIP mixed regression model is then

$$T = S\sqrt{J^{\alpha\alpha}}$$

where $J^{\alpha\alpha}$ is the upper left-hand entry of the inverse information matrix $\mathbf{J}^{-1}$ evaluated at the REML estimates under the null hypothesis. A one-sided test is appropriate because large positive values of T will provide evidence against the null hypothesis. Under H_0, the test statistic T is expected to have an asymptotic standard normal distribution. The finite sample properties of T are investigated by simulation in the next section. Upon confirmation of overdispersion, the alternative ZINB mixed regression model may be considered for fitting the heterogeneous and correlated count data.

4. SAMPLING DISTRIBUTION AND EMPIRICAL POWER

A simulation study is conducted to examine the sampling distribution and the empirical power of the score statistic T under finite sampling situations. The working model under the null hypothesis is taken to be a ZIP mixed regression model with linear predictors:

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 x_{ij} + u_i$$

$$\log\left(\frac{\omega_{ij}}{1-\omega_{ij}}\right) = \gamma_0 + \gamma_1 z_{ij} + v_i$$

for $i = 1, \ldots, m$; $j = 1, \ldots, n$. We set $\beta_0 = 2.5$, $\beta_1 = -1.0$, $\gamma_0 = -1.0$, and $\gamma_1 = 0.5$. Covariates x_{ij} and z_{ij} are generated from a uniform (0,1) distribution whereas the random effects u_i and v_i are assumed to be normally distributed with mean zero and standard deviation 0.2 and 0.1, respectively. In the simulation experiments, $m = 10$, 20 and 40 clusters are used with $n = 10$, 20 and 40 observations per cluster.

The empirical ordered T statistics based on 1000 replications from the above working model are compared with the corresponding quantiles of the standard normal distribution. The Q–Q plots, presented in Figure 1, show that the sampling distribution of T follows closely the asymptotic

DOI: 10.1002/sim

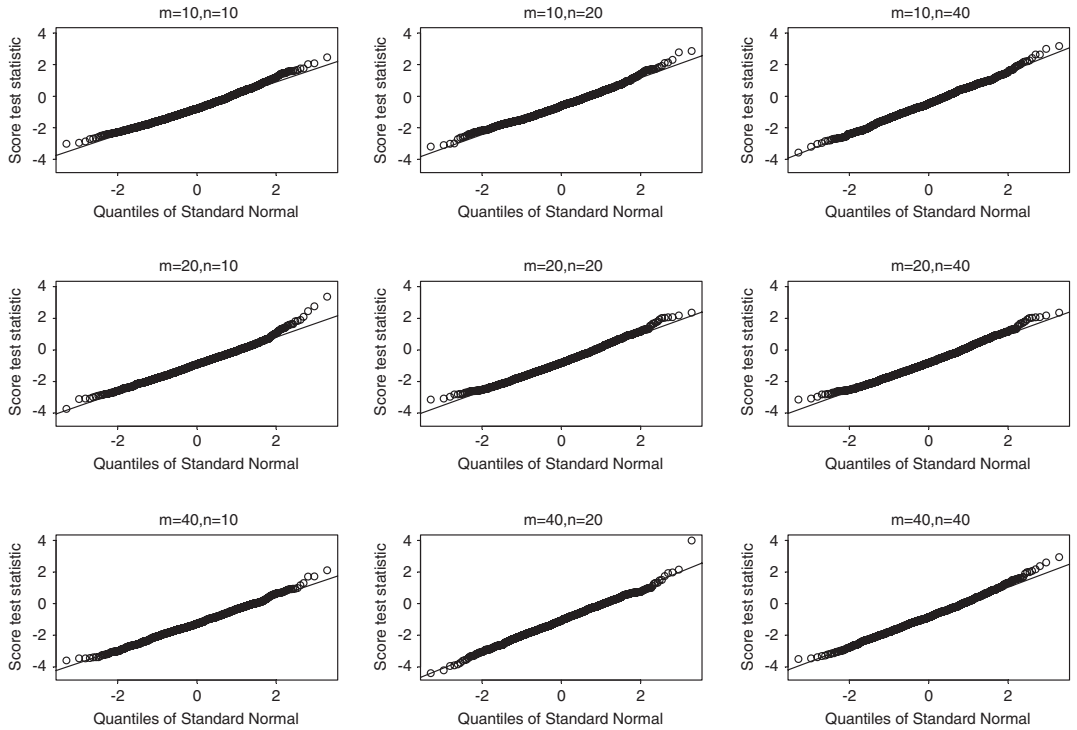

Figure 1. Q–Q plots of ordered score test statistics against standard normal quantiles based on 1000 replications generated from the ZIP mixed regression model.

N(0, 1) distribution for most of the settings chosen. As expected, the approximation improves with more observations per cluster and a larger number of clusters.

We next investigate the power of T for detecting overdispersion. Performance of the test procedure is evaluated under the ZINB mixed regression model, with linear predictors having the same specifications and associated parameter values as defined previously. The empirical power of the test (for given α) is calculated using $Z_{1-\tau}$, which denote the estimated upper tail probabilities of T at the $(1-\tau)$ per cent percentile of N(0, 1) under $H_1 : \alpha > 0$, i.e.

$$P(T > Z_{1-\tau}) \approx \sum_{k=1}^{1000} I[T_k > Z_{1-\tau} | H_1]/1000$$

where T_k is the observed score statistic at the kth replicated trial, $k = 1, \ldots, 1000$. Both small and large degrees of overdispersion, $\alpha = 0.05$, 0.10, 0.20 and 0.40, are considered at nominal significance levels 0.1, 0.05 and 0.01.

The results in Table I demonstrate that the proposed score test is reasonably powerful in rejecting the null hypothesis under the alternative $H_1 : \alpha > 0$. As expected, by increasing the sample size in

DOI: 10.1002/sim

Table I. Empirical power of the score test based on 1000 replications generated from the ZINB mixed regression model.

		Significance level τ			Significance level τ		
m	n	0.10	0.05	0.01	0.10	0.05	0.01
			$\alpha = 0.05$			$\alpha = 0.10$	
10	10	0.237	0.155	0.066	0.645	0.559	0.401
	20	0.341	0.256	0.126	0.879	0.834	0.695
	40	0.879	0.821	0.664	0.996	0.993	0.972
20	10	0.531	0.424	0.259	0.947	0.911	0.823
	20	0.824	0.750	0.578	1.000	1.000	0.999
	40	0.993	0.985	0.934	1.000	1.000	1.000
40	10	0.605	0.511	0.323	0.999	0.998	0.996
	20	0.986	0.967	0.899	1.000	1.000	1.000
	40	1.000	1.000	1.000	1.000	1.000	1.000
			$\alpha = 0.20$			$\alpha = 0.40$	
10	10	0.999	0.997	0.987	1.000	0.999	0.998
	20	1.000	0.997	0.990	1.000	1.000	1.000
	40	1.000	1.000	1.000	1.000	1.000	1.000
20	10	1.000	1.000	0.999	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000
	40	1.000	1.000	1.000	1.000	1.000	1.000
40	10	1.000	1.000	1.000	1.000	1.000	1.000
	20	1.000	1.000	1.000	1.000	1.000	1.000
	40	1.000	1.000	1.000	1.000	1.000	1.000

terms of more clusters or greater number of observations per cluster, a more powerful test can be produced. The empirical power also improves when the degree of overdispersion increases.

5. APPLICATION

We consider the analysis of a set of pancreas disorder inpatient length of stay (LOS) data for a group of 261 patients hospitalized in Western Australia between 1998 and 1999 [20]. Pancreas disorder encompasses acute pancreatitis, chronic pancreatitis and other minor classifications. The empirical LOS frequency distribution exhibits zero-inflation and simultaneous overdispersion, owing to the underlying disease characteristics and available treatment options. The 45 patients (17 per cent) with same-day separations constituted the zero counts, while a few patients who underwent endoscopic surgery sustained prolonged LOS. In addition to LOS, information on clinical- and patient-related characteristics was extracted from the hospital discharge database. Available covariates were: age (in years), gender (0 = female, 1 = male), marital status (0 = married, 1 = single/others), Aboriginality (0 = non-aboriginal, 1 = aboriginal), payment type (0 = public, 1 = private/others), admission status (0 = elective, 1 = emergency), treatment classification (0 = GP/general medicine/gastroenterology, 1 = general surgery), and number of diagnoses. For this sample of patients from 36 public hospitals, their average age was 36 years, 35 per cent were female, 48 per cent were married, while 32 per cent were of aboriginal descent. Although emergency admission accounted for the majority of cases (81 per cent), only 12

DOI: 10.1002/sim

Table II. Comparison of goodness-of-fit statistics between ZIP and ZINB mixed regression models fitted to the pancreas disorder LOS data.

Criteria	ZIP mixed model	ZINB mixed model
Log-likelihood	−491.549	−487.282
AIC	1005.098	996.564
BIC	1044.308	1035.774
Pearson statistic	331.963	261.690
Degree of freedom	237	236

per cent of patients had private medical insurance coverage and 36 per cent involved surgical procedures.

Results of fitting various models, including the ZINB mixed regression model, are given by Yau *et al.* [20]. In this study, hospital effects are treated as random because the 36 hospitals contribute only a subset of the hospital population in Western Australia. The analyses suggested that the ZINB mixed regression model accommodating the inter-hospital variations can lead to substantial improvement in overall goodness-of-fit relative to the unadjusted ZINB model. However, it remains to confirm whether the apparent overdispersion is indeed significant among the non-zero counts. The score test proposed in this paper aims to detect the possible overdispersion among the non-zero counts.

Ignoring random hospital effects, the score test statistic of Ridout *et al.* [19] is 5.56, which indicates that the standard ZIP regression model is unsuitable for these data. Moreover, the proposed score test statistic $T = 4.19$ is very significant (p-value <0.001), providing strong evidence of overdispersion in the ZIP mixed regression model. Results of comparing the ZIP and ZINB mixed regression model fits are summarized in Table II. We found that the ZINB mixed regression model is preferable according to all four goodness-of-fit criteria considered. Furthermore, the scale parameter estimate $\hat{\alpha} = 0.103$ is large relative to its standard error of 0.034, confirming significant overdispersion in these clustered LOS counts with extra zeros.

To further confirm the overdispersion among the non-zero counts, half-normal plots displaying the Pearson residuals *versus* half-normal scores, with simulated envelopes added for the ZIP and ZINB mixed regression models [23] are given in Figure 2. It is clear from the plots that the Pearson residuals are lying within the envelope for ZINB but not for ZIP mixed regression model, further confirming the presence of overdispersion among the non-zero.

6. CONCLUDING REMARKS

A score statistic is proposed for testing overdispersion in correlated count data with extra zeros. Unlike the likelihood ratio test, the advantage of the score statistic lies in its computational convenience, as it does not require the more complex ZINB mixed regression model to be fitted. The procedure has been implemented as an S-Plus macro available from the authors. Although the numerical example on pancreas disorder LOS is concerned with clustered data, the score test procedure is also applicable to longitudinal count data, where repeated measures of the variables of interest are collected over time.

DOI: 10.1002/sim

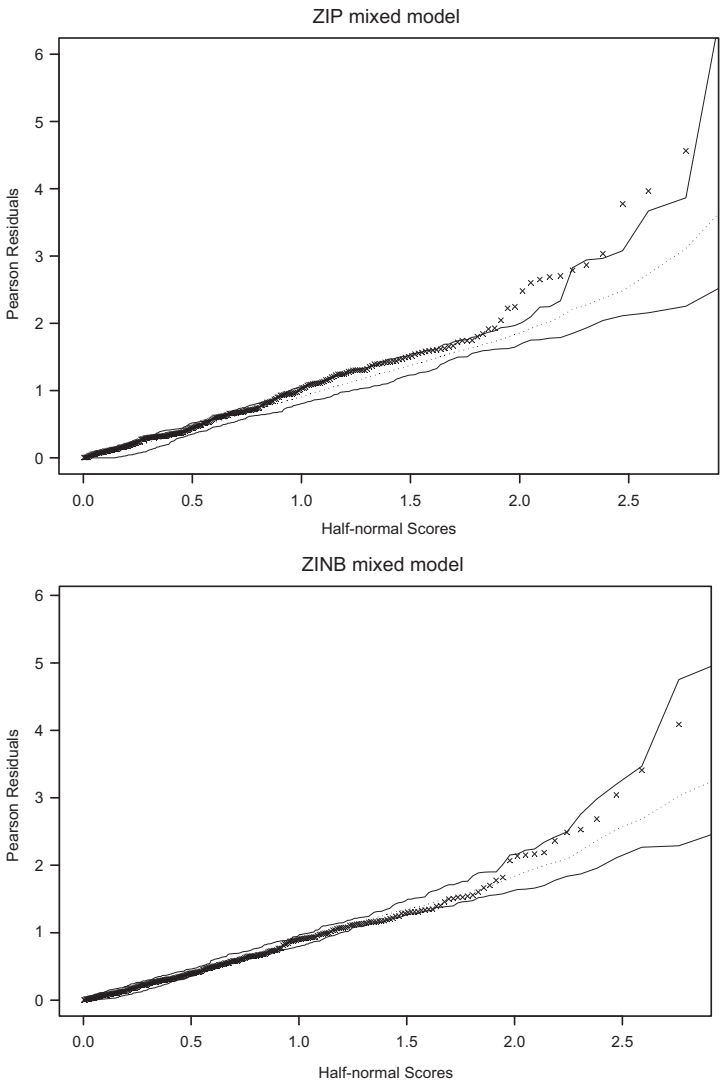

Figure 2. Half-normal plots of Pearson residuals *versus* half-normal scores, with simulated envelopes added, for the ZIP and ZINB mixed regression models fitted to the pancreas disorder LOS data.

Despite the paucity of asymptotic inferences for mixed regression models, results of our simulation study show that the proposed test has high power and the score statistic follows approximately a standard normal distribution, even in small sample settings. From a practical viewpoint, the nominal significance level of the observed score statistic enables the assessment of overdispersion for correlated count data. In the presence of simultaneous zero-inflation and overdispersion, the fit of the alternative ZINB mixed regression model is recommended, and comparisons should be made with the corresponding ZIP counterpart. On the other hand, if the score test gives no indication

DOI: 10.1002/sim

of lack of fit, inferences based on the ZIP mixed regression model can be made with increased confidence, analogous to the independent data situation.

APPENDIX A

The derivation of the efficient score S and the Fisher information matrix $\mathbf{J}$ in Section 3 is given here, which is in principle similar to that presented in Reference [22].

Note that

$$\frac{\Gamma(y+1/\alpha)}{\Gamma(1/\alpha)} = \alpha^{-y} \prod_{k=1}^{y} (\alpha y - \alpha k + 1)$$

We write the fixed effect contribution of the log-likelihood l as $l_1 = \sum_{i,j}(l_{1ij,1} + l_{1ij,2})$, where $l_{1ij,1} = I_{\{y_{ij}=0\}} \log[\omega_{ij} + (1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}]$ and $l_{1ij,2} = I_{\{y_{ij}>0\}}[\log(1-\omega_{ij}) - \log(y_{ij}!) + \sum_{k=1}^{y_{ij}} \log(\alpha y_{ij} - \alpha k + 1) - (y_{ij} + 1/\alpha)\log(1+\alpha\lambda_{ij}) + y_{ij}\log(\lambda_{ij})]$. Then,

$$\frac{\partial l_{1ij,1}}{\partial \alpha} = I_{\{y_{ij}=0\}} \frac{(1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}}{\omega_{ij} + (1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}} \left[\frac{\log(1+\alpha\lambda_{ij})}{\alpha^2} - \frac{\lambda_{ij}}{\alpha(1+\alpha\lambda_{ij})} \right]$$

and

$$\frac{\partial l_{1ij,2}}{\partial \alpha} = I_{\{y_{ij}>0\}} \left[\sum_{k} \frac{y_{ij}-k}{\alpha y_{ij} - \alpha k + 1} + \frac{\log(1+\alpha\lambda_{ij})}{\alpha^2} - \frac{(y_{ij}+1/\alpha)\lambda_{ij}}{1+\alpha\lambda_{ij}} \right]$$

Let $p_{0,ij} = P(y=0|H_0) = \omega_{ij} + (1-\omega_{ij})\mathrm{e}^{-\lambda_{ij}}$. Note that $\lim_{\alpha\to 0}(1+\alpha\lambda_{ij})^{-1/\alpha} = \mathrm{e}^{-\lambda_{ij}}$ and

$$\lim_{\alpha\to 0} \left[\frac{\log(1+\alpha\lambda_{ij})}{\alpha^2} - \frac{\lambda_{ij}}{\alpha(1+\alpha\lambda_{ij})} \right] = \frac{\lambda_{ij}^2}{2}$$

using L'Hospital's rule. It follows that the first-order derivative of the log-likelihood l with respect to α and evaluated under the null hypothesis

$$\left.\frac{\partial l}{\partial \alpha}\right|_{\alpha=0} = \sum_{i,j} \left.\frac{\partial l_{1ij,1}}{\partial \alpha}\right|_{\alpha=0} + \sum_{i,j} \left.\frac{\partial l_{1ij,2}}{\partial \alpha}\right|_{\alpha=0} = \frac{1}{2}\sum_{i,j} \left\{ [(y_{ij}-\lambda_{ij})^2 - y_{ij}] - I_{\{y_{ij}=0\}} \lambda_{ij}^2 \frac{\omega_{ij}}{p_{0,ij}} \right\}$$

which gives the efficient score.

Next, to obtain elements of the Fisher information matrix $\mathbf{J}$, the expected negative second-order derivatives of l with respect to any two of parameters α, β, γ, u and v are derived as follows. Similar to the first-order derivatives above, L'Hospital's rule is frequently applied in evaluating the quantities under the null hypothesis.

$$\frac{\partial^2 l_{1ij,1}}{\partial \alpha^2} = \frac{I_{\{y_{ij}=0\}}(1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}}{[\omega_{ij} + (1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha}]^2}$$
$$\times \left[(1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha} \left(\frac{\log(1+\alpha\lambda_{ij})}{\alpha^2} - \frac{\lambda_{ij}}{1+\alpha\lambda_{ij}} \right) \right.$$

DOI: 10.1002/sim

$$-(\omega_{ij}+(1-\omega_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha})\left\{\left(\frac{\log(1+\alpha\lambda_{ij})}{\alpha^2}-\frac{\lambda_{ij}}{1+\alpha\lambda_{ij}}\right)^2\right.$$

$$\left.-\frac{2\log(1+\alpha\lambda_{ij})}{\alpha^3}+\frac{\lambda_{ij}(2+3\alpha\lambda_{ij})}{\alpha^2(1+\alpha\lambda_{ij})^2}\right\}\Bigg]$$

$$\frac{\partial^2 l_{1ij,2}}{\partial\alpha^2}=I_{\{y_{ij}>0\}}\left[\sum_{k=1}^{y_{ij}}\frac{-(y_{ij}-k)^2}{(\alpha y_{ij}-\alpha k+1)^2}-\frac{2\log(1+\alpha\lambda_{ij})}{\alpha^3}\right.$$

$$\left.+\frac{2\lambda_{ij}}{\alpha^2(1+\alpha\lambda_{ij})}+\frac{(y_{ij}+1/\alpha)\lambda_{ij}^2}{(1+\alpha\lambda_{ij})^2}\right]$$

$$\left.\frac{\partial^2 l_{1ij,1}}{\partial\alpha\partial\lambda_{ij}}\right|_{\alpha=0}=\frac{I_{\{y_{ij}=0\}}}{p_{0,ij}}\left[-\frac{\lambda_{ij}^2}{2}(1-\omega_{ij})\mathrm{e}^{-\lambda_{ij}}+(1-\omega_{ij})\lambda_{ij}\mathrm{e}^{-\lambda_{ij}}+\frac{\lambda_{ij}^2}{2p_{0,ij}}(1-\omega_{ij})^2\mathrm{e}^{-2\lambda_{ij}}\right]$$

$$\left.\frac{\partial^2 l_{1ij,2}}{\partial\alpha\partial\lambda_{ij}}\right|_{\alpha=0}=\frac{I_{\{y_{ij}>0\}}(\lambda_{ij}-y_{ij})}{(1+\alpha\lambda_{ij})^2},\quad \left.\frac{\partial^2 l_{1ij,1}}{\partial\alpha\partial\omega_{ij}}\right|_{\alpha=0}=-\frac{I_{\{y_{ij}=0\}}}{p_{0,ij}^2}\frac{\lambda_{ij}^2}{2}\mathrm{e}^{-\lambda_{ij}},\quad \frac{\partial^2 l_{1ij,2}}{\partial\alpha\partial\omega_{ij}}=0$$

$$\left.\frac{\partial^2 l_{1ij,1}}{\partial\lambda_{ij}^2}\right|_{\alpha=0}=\frac{I_{\{y_{ij}=0\}}}{p_{0,ij}}\left[(1-\omega_{ij})\mathrm{e}^{-\lambda_{ij}}-\frac{1}{p_{0,ij}}(1-\omega_{ij})^2\mathrm{e}^{-2\lambda_{ij}}\right]$$

$$\left.\frac{\partial^2 l_{1ij,2}}{\partial\lambda_{ij}^2}\right|_{\alpha=0}=-I_{\{y_{ij}>0\}}\frac{y_{ij}}{\lambda_{ij}^2},\quad \left.\frac{\partial^2 l_{1ij,1}}{\partial\omega_{ij}^2}\right|_{\alpha=0}=-\frac{I_{\{y_{ij}=0\}}}{p_{0,ij}^2}(1-\mathrm{e}^{-\lambda_{ij}})^2$$

$$\left.\frac{\partial^2 l_{1ij,2}}{\partial\omega_{ij}^2}\right|_{\alpha=0}=-\frac{I_{\{y_{ij}>0\}}}{(1-\omega_{ij})^2},\quad \left.\frac{\partial^2 l_{1ij,1}}{\partial\lambda_{ij}\partial\omega_{ij}}\right|_{\alpha=0}=\frac{I_{\{y_{ij}=0\}}\mathrm{e}^{-\lambda_{ij}}}{p_{0,ij}^2}$$

and

$$\frac{\partial^2 l_{1ij,2}}{\partial\lambda_{ij}\partial\omega_{ij}}=0$$

Since the response variable y_{ij} follows a ZIP distribution under $H_0:\alpha=0$, the first three moments of y_{ij} conditional on random effects are given by:

$$E(y_{ij})=(1-\omega_{ij})\lambda_{ij},\quad E(y_{ij}^2)=(1-\omega_{ij})\lambda_{ij}(1+\lambda_{ij}),\quad E(y_{ij}^3)=(1-\omega_{ij})\lambda_{ij}(1+3\lambda_{ij}+\lambda_{ij}^3)$$

Also note that $E(I_{\{y_{ij}>0\}})=(1-\omega_{ij})(1-\mathrm{e}^{-\lambda_{ij}})$, $E(I_{\{y_{ij}=0\}})=p_{0,ij}$, $\partial\lambda_{ij}/\partial\beta=\lambda_{ij}\mathbf{x}_{ij}$, $\partial\lambda_{ij}/\partial u=\lambda_{ij}\mathbf{w}_{ij}$, $\partial\omega_{ij}/\partial\gamma=\omega_{ij}(1-\omega_{ij})\mathbf{z}_{ij}$ and $\partial\omega_{ij}/\partial v=\omega_{ij}(1-\omega_{ij})\mathbf{w}_{ij}$. Expectations of negative

DOI: 10.1002/sim

second-order derivatives of l taken under $H_0 : \alpha = 0$ conditional on random effects are therefore obtained:

$$E\left(-\frac{\partial^2 l}{\partial \alpha^2}\right) = E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,1}}{\partial \alpha^2}\right) + E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,2}}{\partial \alpha^2}\right)$$

$$= \sum_{i,j}\left[\frac{1}{2}(1-\omega_{ij})\lambda_{ij}^2 - \frac{\lambda_{ij}^4}{4}\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\right]$$

$$E\left(-\frac{\partial^2 l}{\partial \alpha \partial \beta}\right) = E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,1}}{\partial \alpha \partial \lambda_{ij}}\frac{\partial \lambda_{ij}}{\partial \beta}\right) + E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,2}}{\partial \alpha \partial \lambda_{ij}}\frac{\partial \lambda_{ij}}{\partial \beta}\right) = \sum_{i,j}\frac{\lambda_{ij}^3}{2}\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\mathbf{x}_{ij}$$

$$E\left(-\frac{\partial^2 l}{\partial \alpha \partial u}\right) = E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,1}}{\partial \alpha \partial \lambda_{ij}}\frac{\partial \lambda_{ij}}{\partial u}\right) + E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,2}}{\partial \alpha \partial \lambda_{ij}}\frac{\partial \lambda_{ij}}{\partial u}\right) = \sum_{i,j}\frac{\lambda_{ij}^3}{2}\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\mathbf{w}_{ij}$$

$$E\left(-\frac{\partial^2 l}{\partial \alpha \partial \gamma}\right) = E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,1}}{\partial \alpha \partial \omega_{ij}}\frac{\partial \omega_{ij}}{\partial \gamma}\right) = \sum_{i,j}\frac{\lambda_{ij}^2}{2}\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\mathbf{z}_{ij}$$

$$E\left(-\frac{\partial^2 l}{\partial \alpha \partial v}\right) = E\left(-\sum_{i,j}\frac{\partial^2 l_{1ij,1}}{\partial \alpha \partial \omega_{ij}}\frac{\partial \omega_{ij}}{\partial v}\right) = \sum_{i,j}\frac{\lambda_{ij}^2}{2}\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\mathbf{w}_{ij}$$

$$E\left(-\frac{\partial^2 l}{\partial \beta \partial \beta'}\right) = E\left(-\sum_{i,j}\frac{\partial \lambda_{ij}}{\partial \beta}\left(\frac{\partial^2 l_{1ij,1}}{\partial \lambda_{ij}^2} + \frac{\partial^2 l_{1ij,2}}{\partial \lambda_{ij}^2}\right)\frac{\partial \lambda_{ij}}{\partial \beta'}\right)$$

$$= \sum_{i,j}\left[\lambda_{ij}(1-\omega_{ij}) - \lambda_{ij}^2\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\right]\mathbf{x}_{ij}\mathbf{x}_{ij}'$$

$$E\left(-\frac{\partial^2 l}{\partial \beta \partial \gamma'}\right) = E\left(-\sum_{i,j}\frac{\partial \lambda_{ij}}{\partial \beta}\left(\frac{\partial^2 l_{1ij,1}}{\partial \lambda_{ij}\partial \omega_{ij}} + \frac{\partial^2 l_{1ij,2}}{\partial \lambda_{ij}\partial \omega_{ij}}\right)\frac{\partial \omega_{ij}}{\partial \gamma'}\right)$$

$$= \sum_{i,j}\left[-\lambda_{ij}\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\right]\mathbf{x}_{ij}\mathbf{z}_{ij}'$$

$$E\left(-\frac{\partial^2 l}{\partial \gamma \partial \gamma'}\right) = E\left(-\sum_{i,j}\frac{\partial \omega_{ij}}{\partial \gamma}\left(\frac{\partial^2 l_{1ij,1}}{\partial \omega_{ij}^2} + \frac{\partial^2 l_{1ij,2}}{\partial \omega_{ij}^2}\right)\frac{\partial \omega_{ij}}{\partial \gamma'}\right) = \sum_{i,j}\omega_{ij}^2\left(\frac{1}{p_{0,ij}} - 1\right)\mathbf{z}_{ij}\mathbf{z}_{ij}'$$

$$E\left(-\frac{\partial^2 l}{\partial u \partial u'}\right) = E\left(-\sum_{i,j}\frac{\partial \lambda_{ij}}{\partial u}\left(\frac{\partial^2 l_{1ij,1}}{\partial \lambda_{ij}^2} + \frac{\partial^2 l_{1ij,2}}{\partial \lambda_{ij}^2}\right)\frac{\partial \lambda_{ij}}{\partial u'} - \frac{\partial^2 l_2}{\partial u \partial u'}\right)$$

$$= \sum_{i,j}\left[\lambda_{ij}(1-\omega_{ij}) - \lambda_{ij}^2\omega_{ij}\left(1-\frac{\omega_{ij}}{p_{0,ij}}\right)\right]\mathbf{w}_{ij}\mathbf{w}_{ij}' + \sigma_u^{-2}\mathbf{I}_m$$

DOI: 10.1002/sim

$$E\left(-\frac{\partial^2 l}{\partial v \partial v'}\right) = E\left(-\sum_{i,j}\frac{\partial \omega_{ij}}{\partial v}\left(\frac{\partial^2 l_{1ij,1}}{\partial \omega_{ij}^2} + \frac{\partial^2 l_{1ij,2}}{\partial \omega_{ij}^2}\right)\frac{\partial \omega_{ij}}{\partial v'} - \frac{\partial^2 l_2}{\partial v \partial v'}\right)$$

$$= \sum_{i,j}\omega_{ij}^2\left(\frac{1}{p_{0,ij}} - 1\right)\mathbf{w}_{ij}\mathbf{w}_{ij}' + \sigma_v^{-2}\mathbf{I}_m$$

where $\mathbf{I}_m$ denotes an $m \times m$ identity matrix.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewer for helpful comments. This research is supported by grants from the Australian Research Council and the Research Grants Council of Hong Kong. The authors are grateful to the Health Information Centre, Department of Health, Western Australia, for provision of the pancreas disorder length of stay data.

REFERENCES

1. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**:1–14.
2. 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*, *Series A* 1999; **162**:195–209.
3. Dietz E, Böhning D. On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis* 2000; **34**:441–459.
4. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 2000; **56**:1030–1039.
5. Yau KKW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine* 2001; **20**:2907–2920.
6. 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.
7. 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.
8. Lee AH, Wang K, Scott JA, Yau KKW, McLachlan GJ. Multilevel zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research* 2006; **15**:47–61.
9. McLachlan GJ. On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research* 1997; **6**:76–98.
10. Dobbie MJ, Welsh AH. Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics* 2001; **43**:431–444.
11. Hall DB, Zhang Z. Marginal models for zero inflated clustered data. *Statistical Modelling* 2004; **4**:161–180.
12. Van den Broek J. A score test for zero-inflation in a Poisson distribution. *Biometrics* 1995; **51**:738–743.
13. Lee AH, Wang K, Yau KKW. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* 2001; **43**:963–975.
14. Deng D, Paul SR. Score tests for zero inflation in generalized linear models. *The Canadian Journal of Statistics* 2000; **27**:563–570.
15. Jansakul N, Hinde JP. Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis* 2002; **40**:75–96.
16. Ugarte MD, Ibáñez B, Militino AF. Testing for Poisson zero inflation in disease mapping. *Biometrical Journal* 2004; **46**:526–539.
17. Lee AH, Xiang L, Fung WK. Sensitivity of score tests for zero-inflation in count data. *Statistics in Medicine* 2004; **23**:2757–2769.
18. Xiang L, Lee AH, Yau KKW, McLachlan GJ. A score test for zero-inflation in correlated count data. *Statistics in Medicine* 2006; **25**:1660–1671.

DOI: 10.1002/sim

19. Ridout M, Hinde J, Demtrio CGB. A score test for testing zero inflated Poisson regression model against zero inflated negative binomial alternatives. *Biometrics* 2001; **57**:219–223.
20. 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**:437–452.
21. McGilchrist CA. Estimation in generalized linear mixed models. *Journal of the Royal Statistical Society*, *Series B* 1994; **56**:61–69.
22. Deng D, Paul SR. Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica* 2005; **15**:257–276.
23. Vieira AMC, Hinde JP, Demtrio CGB. Zero-inflated proportion data models applied to biological control assay. *Journal of Applied Statistics* 2000; **27**:373–389.

DOI: 10.1002/sim