STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:1660–1671
Published online 5 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2308

A score test for zero-inflation in correlated count data

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

[1] *Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, G.P.O. Box U 1987, Perth, WA 6845, Australia*
[2] *Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon Tong, Hong Kong*
[3] *Department of Mathematics, University of Queensland, St. Lucia, Brisbane, Queensland, 4072, Australia*

SUMMARY

To account for the preponderance of zero counts and simultaneous correlation of observations, a class of zero-inflated Poisson mixed regression models is applicable for accommodating the within-cluster dependence. In this paper, a score test for zero-inflation is developed for assessing correlated count data with excess zeros. The sampling distribution and the 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 further illustrated using a data set on recurrent urinary tract infections. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: count data; Poisson mixed model; power; score test; zero-inflation

1. INTRODUCTION

Count data with excess zeros relative to a Poisson distribution are commonly encountered in many biomedical and public health applications. Failure to account for the extra zeros may result in biased parameter estimates and misleading inferences [1]. The zero-inflated Poisson (ZIP) model, which mixes the Poisson distribution with a degenerate component of point mass at zero, has become a popular approach to analyse such data. Böhning [2] reviewed the related literature and provided a variety of examples from different disciplines. Further applications of the ZIP regression model can be found in References [3–5].

[*] Correspondence to: Andy H. Lee, Department of Epidemiology and Biostatistics, School of Public Health, Curtin University of Technology, G.P.O. Box U 1987, Perth, WA 6845, Australia.
[†] E-mail: andy.lee@curtin.edu.au

Contract/grant sponsor: Australian Research Council
Contract/grant sponsor: National Health and Medical Research Council of Australia
Contract/grant sponsor: Research Grants Council of Hong Kong

Received December 2004
Accepted May 2005

Often, due to the hierarchical study design or the data collection procedure (such as longitudinal repeated measures), zero-inflation and lack of independence may be present simultaneously as a consequence of the inherent correlation structure and underlying heterogeneity. To adjust for the dependency of observations, Yau and Lee [6] introduced extra random components to the ZIP regression model. Wang *et al.* [7] further developed a class of ZIP mixed models where random effects are incorporated within the linear predictors of both the Poisson part and the zero-mixing probability, using the generalized linear mixed modelling approach of McGilchrist [8]. Parameter estimation is facilitated via an EM algorithm.

In applications to potentially zero-inflated count data, it is important to assess whether the ZIP model assumption is indeed appropriate. A score test for zero-inflation was first proposed by Van den Broek [9]. Lee *et al.* [5] compared the score statistic with its likelihood ratio and Wald counterparts in a simulation study. Recently, these tests have been extended from the standard Poisson to other settings, including discrete generalized linear models [10] and zero-inflated negative binomial (ZINB) [11]. Jansakul and Hinde [12] modified the score test to the general situation where the zero-mixing probability is allowed to depend on covariates.

The aim of this paper is to develop an appropriate test to assess correlated count data with an apparently high frequency of zeros. A score test for zero-inflation is proposed for the Poisson mixed regression model, in the manner of Van den Broek [9]. After briefly reviewing the ZIP mixed model in Section 2, the zero-inflation hypothesis and corresponding score test are specified in Section 3. In Section 4, the sampling distribution of the score test statistic and its power properties are investigated by a simulation study. In the presence of extra zeros, it is expected that the assessment of zero-inflation will enable practitioners to draw valid inferences on count data models. An example arising from the analysis of recurrent urinary tract infections (UTI) in elderly women, where the correlated data collected from a retrospective cohort study exhibit a preponderance of zero counts, is used to illustrate the test procedure. Following the illustrative example in Section 5, some discussions on further generalizations of the test procedure are given in Section 6.

2. ZIP MIXED MODEL

Suppose a discrete response variable Y follows a ZIP distribution defined by

$$\begin{aligned} P\{Y=0\} &= \phi + (1-\phi) f(0;\eta) \\ P\{Y=y\} &= (1-\phi) f(y;\eta), \quad y>0 \end{aligned} \tag{1}$$

where $f(0;\eta)=\exp(-\exp(\eta))$

$$f(y;\eta)=\exp\{y\eta-\exp(\eta)-\log(y!)\}, \quad y>0 \tag{2}$$

and $-(f(0;\eta)/(1-f(0;\eta)))\leqslant\phi<1$ so that the ZIP distribution allows more zeros than those permitted by the Poisson ($\phi=0$), while $\phi<0$ corresponds to the zero-deflated situation [13]. Denoting $\mu=\exp(\eta)$, it can be shown that $E(Y)=(1-\phi)\mu$, and $\mathrm{var}(Y)=E(Y)(1+\mu-E(Y))$.

Consider a two-level hierarchical setting where y_{ij} ($i=1,\dots,m$, $j=1,\dots,n_i$) represents the jth response within the ith cluster and let $N=\sum_i n_i$. It may be assumed that the observations are independent between clusters but within-cluster correlation is anticipated. The ZIP mixed regression model assumes both μ_{ij} and ϕ_{ij} to be separate functions of some covariates and random effects to adjust for the clustering [7]. Under this model framework,

$$\begin{aligned} g(\mu_{ij}) &= X_{ij}'\beta + u_i \\ h(\phi_{ij}) &= W_{ij}'\alpha + v_i \end{aligned} \tag{3}$$

where X_{ij} and W_{ij} are vectors of covariates, β and α are the corresponding $q \times 1$ and $r \times 1$ vectors of regression coefficients. The random components u_i and v_i are independent and have $\mathrm{N}(0, \sigma_u^2)$ and $\mathrm{N}(0, \sigma_v^2)$ distributions, respectively, while $g(\cdot)$ and $h(\cdot)$ are known as link functions. For simplicity of presentation but without loss of generality, a log-link is adopted for the Poisson part and the zero mixing proportion is held fixed, i.e. $h(\phi_{ij}) = \phi$, in subsequent derivations.

Parameter estimation can be achieved following the restricted maximum likelihood (REML) approach of McGilchrist [8]. The best linear unbiased prediction-type log-likelihood is given by $l = l_1 + l_2$, where

$$l_1 = \sum_{i,j} \{ I_{(y_{ij}=0)} \log[\phi + (1-\phi) f(0, \eta_{ij})] + I_{(y_{ij}>0)} \log[(1-\phi) f(y_{ij}, \eta_{ij})] \}$$

with indicator function $I_{(.)}$ taking the value 1 when satisfying the specified condition and 0 otherwise, and $l_2 = -\frac{1}{2}[m \log(2\pi\sigma_u^2) + \sigma_u^{-2} \sum_i u_i^2]$, with m denoting the number of clusters. Here, l can be viewed as a penalized log-likelihood function with l_2 being the penalty for the conditional log-likelihood l_1 when the random effects are conditionally fixed. With suitable initial values, the REML estimates $\hat{\beta}$ and $\hat{u}$ (for conditionally fixed u) can be obtained by maximizing the log-likelihood l via a numerical procedure, such as the extended version of the EM algorithm for overdispersed count data [14, 15]. The estimate of variance component σ_u^2 is then computed from an estimating equation involving $\hat{\beta}$ and $\hat{u}$, details of which are described in Reference [7].

3. THE SCORE TEST FOR ZERO-INFLATION

The score function of the parameters is derived as follows. Let $\gamma = \phi/(1-\phi)$, $-f(0;\eta) \leqslant \gamma < \infty$ for $-(f(0;\eta)/(1-f(0;\eta))) \leqslant \phi < 1$. Testing the null hypothesis $H_0 : \phi = 0$ against $H_1 : \phi \neq 0$ is equivalent to testing $H_0^* : \gamma = 0$ against $H_1^* : \gamma \neq 0$. Note that the conditional log-likelihood l_1 can be rewritten as $l_1(\gamma, \eta; y) = \sum_{i,j} l_{1ij}$, where

$$l_{1ij} = -\log(1+\gamma) + I_{(y_{ij}=0)} \log[\gamma + \exp(-\exp(\eta_{ij}))] + I_{(y_{ij}>0)}[\eta_{ij} y_{ij} - \exp(\eta_{ij}) + \log(y_{ij}!)]$$

Let $\tau = \sigma_u^2$. Taking the first and second derivatives of l with respect to β, u, τ and γ, the score function $U(\beta, u, \tau, \gamma)$ and the Fisher information matrix $\mathfrak{I}(\beta, u, \tau, \gamma)$ can be obtained. Specifically, the first derivatives are given by

$$\frac{\partial l}{\partial \beta} = \sum_{i,j} \frac{\partial l_{1ij}}{\partial \beta}, \quad \frac{\partial l}{\partial u} = \sum_{i,j} \frac{\partial l_{1ij}}{\partial u} + \frac{\partial l_2}{\partial u}, \quad \frac{\partial l}{\partial \tau} = -\frac{m}{2\tau} + \frac{1}{2\tau^2} \sum_i u_i^2$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i,j} \frac{\partial l_{1ij}}{\partial \gamma} = \sum_{i,j} \left[-\frac{1}{1+\gamma} + I_{(y_{ij}=0)} \frac{1}{\gamma + f(0, \eta_{ij})} \right]$$

Under the null hypothesis $H_0^*: \gamma = 0$, the reduced model is the Poisson mixed model by setting $\phi = 0$ in l_1. Suppose $\tilde{\beta}$, $\tilde{u}$ and $\tilde{\tau}$ denote the corresponding REML parameter estimates of the Poisson mixed regression model. The score function is

$$U(\tilde{\beta}, \tilde{u}, \tilde{\tau}, 0) = \left(0, \ldots, 0, \sum_{i,j} \left[\frac{I_{(y_{ij}=0)}}{f(0, \tilde{\eta}_{ij})} - 1\right]\right)'$$

It can be shown that

$$\frac{\partial l_{1ij}}{\partial \eta_{ij}} = I_{(y_{ij}=0)} \frac{-f(0,\eta_{ij})\exp(\eta_{ij})}{\gamma + f(0,\eta_{ij})} + I_{(y_{ij}>0)}[y_{ij} - \exp(\eta_{ij})]$$

$$\frac{\partial^2 l_{1ij}}{\partial \eta_{ij}^2} = I_{(y_{ij}=0)} \left\{ \frac{f(0,\eta_{ij})[\exp(\eta_{ij})]^2}{\gamma + f(0,\eta_{ij})} + \frac{-f^2(0,\eta_{ij})[\exp(\eta_{ij})]^2}{[\gamma + f(0,\eta_{ij})]^2} + \frac{-f(0,\eta_{ij})\exp(\eta_{ij})}{\gamma + f(0,\eta_{ij})} \right\}$$
$$+ I_{(y_{ij}>0)}[-\exp(\eta_{ij})]$$

and

$$\frac{\partial^2 l_{1ij}}{\partial \eta_{ij} \partial \gamma} = I_{(y_{ij}=0)} \frac{f(0,\eta_{ij})\exp(\eta_{ij})}{[\gamma + f(0,\eta_{ij})]^2}$$

The expected Fisher information matrix is then

$$\Im(\beta, u, \tau, \gamma) = \begin{pmatrix} \Im_{\beta\beta} & \Im_{\beta u} & \Im_{\beta\tau} & \Im_{\beta\gamma} \\ & \Im_{uu} & \Im_{u\tau} & \Im_{u\gamma} \\ & & \Im_{\tau\tau} & \Im_{\tau\gamma} \\ & & & \Im_{\gamma\gamma} \end{pmatrix} \qquad (4)$$

where entries of $\Im(\beta, u, \tau, \gamma)$ under H_0^* are obtained by evaluating the second derivatives of l at $\gamma = 0$, the formula of which are given in Appendix A. The matrix $\Im(\beta, u, \tau, \gamma)$ may be partitioned as follows:

$$\begin{pmatrix} \Im_{11} & \Im_{12} \\ \Im'_{12} & \Im_{22} \end{pmatrix} \qquad (5)$$

where $\Im_{22} = \Im_{\gamma\gamma} = \sum_{i,j}[(1/f(0;\eta_{ij})) - 1]$, $\Im'_{12} = (1'_N BT, 1'_N BP, 0)$, and

$$\Im_{11} = \begin{pmatrix} -T'BT & -T'BP & 0 \\ -P'BT & -P'BP - \tau^{-1}I & -\tau^{-2}u \\ 0 & -\tau^{-2}u' & -\tau^{-2}m/2 + \tau^{-3}u'u \end{pmatrix}$$

with matrices T, P and B defined in Appendix A. Furthermore, let $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ be the inverse of $\Im$ having the same partition as (5). In view of the structure of the score function U, only Σ_{22} will appear in the expression of the score test statistic, with $\Sigma_{22}^{-1} = \Im_{22} - \Im_{12}' \Im_{11}^{-1} \Im_{12}$. Consequently, the score statistic for testing $H_0^* : \gamma = 0$ is given by

$$\begin{aligned} S(\tilde{\beta}, \tilde{u}, \tilde{\tau}, 0) &= U'(\tilde{\beta}, \tilde{u}, \tilde{\tau}, 0) \tilde{\Im}^{-1} U(\tilde{\beta}, \tilde{u}, \tilde{\tau}, 0) \\ &= \frac{\sum_{i,j} [(I_{(y_{ij}=0)} / f(0; \tilde{\eta}_{ij})) - 1]^2}{\sum_{i,j} [(1/f(0; \tilde{\eta}_{ij})) - 1] - \tilde{\Im}_{12}' \tilde{\Im}_{11}^{-1} \tilde{\Im}_{12}} \end{aligned} \tag{6}$$

where $\tilde{\eta}_{ij}$ and $\tilde{\Im}$'s are evaluated at the REML estimates $(\tilde{\beta}, \tilde{u}, \tilde{\tau})$.

From (6), the score statistic S exhibits a quadratic form so standard statistical theory [16] implies that S has an asymptotic χ_1^2 distribution under the null hypothesis H_0^*. For the simple case when $u = 0$, S reduces to the ordinary score test statistic of Van den Broek [9]. The finite sample properties of the score test statistic will be examined by simulation in the next section.

4. SAMPLING DISTRIBUTION AND POWER INVESTIGATION

4.1. Sampling distribution

A simulation study is conducted to investigate the distribution of the score statistic under finite sample situations. The working model under the null hypothesis is taken to be a Poisson mixed regression model with linear predictor

$$\log(\mu_{ij}) = a + bx_{ij} + u_i \tag{7}$$

for $i = 1, \ldots, m$; $j = 1, \ldots, n$. Following the simulation design of Van den Broek [9], we set $a = 0.5$ and $b = 1$. The single covariate x_{ij} is generated from a uniform $(0, 1)$ distribution whereas the random effect u_i is assumed to follow a normal distribution with mean zero and variance 0.25. Therefore, for given x_{ij} the approximate 95 per cent confidence limits for the mean response of y_{ij} are 0.6 and 12. We consider $m = 5$, 10 and 20 clusters with $n = 10$, 20 and 40 observations per cluster.

The empirical ordered S statistics based on 1000 replications from model (7) are first compared with the corresponding quantiles of the χ_1^2 distribution. The *Q–Q* plots, presented in Figure 1, show that the sampling distribution of S follows closely the asymptotic χ_1^2 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 assess the effect of varying σ_u on the sampling distribution of S. Again 1000 replications are generated for $\sigma_u = 0.1$, 0.5 and 1 on two sample-size combinations: $(m = 5,\ n = 10)$ and $(m = 20,\ n = 20)$. The resulting *Q–Q* plots, given in Figure 2, confirm that the asymptotic χ_1^2 distribution is reasonable for smaller σ_u value, i.e. reduced variation in the random component. The asymptotic null distribution works less well for larger σ_u because it would lead to more variations on the parameter estimates and consequently a larger variance for the score statistic. Moreover, an increase in sample size also leads to a better agreement.

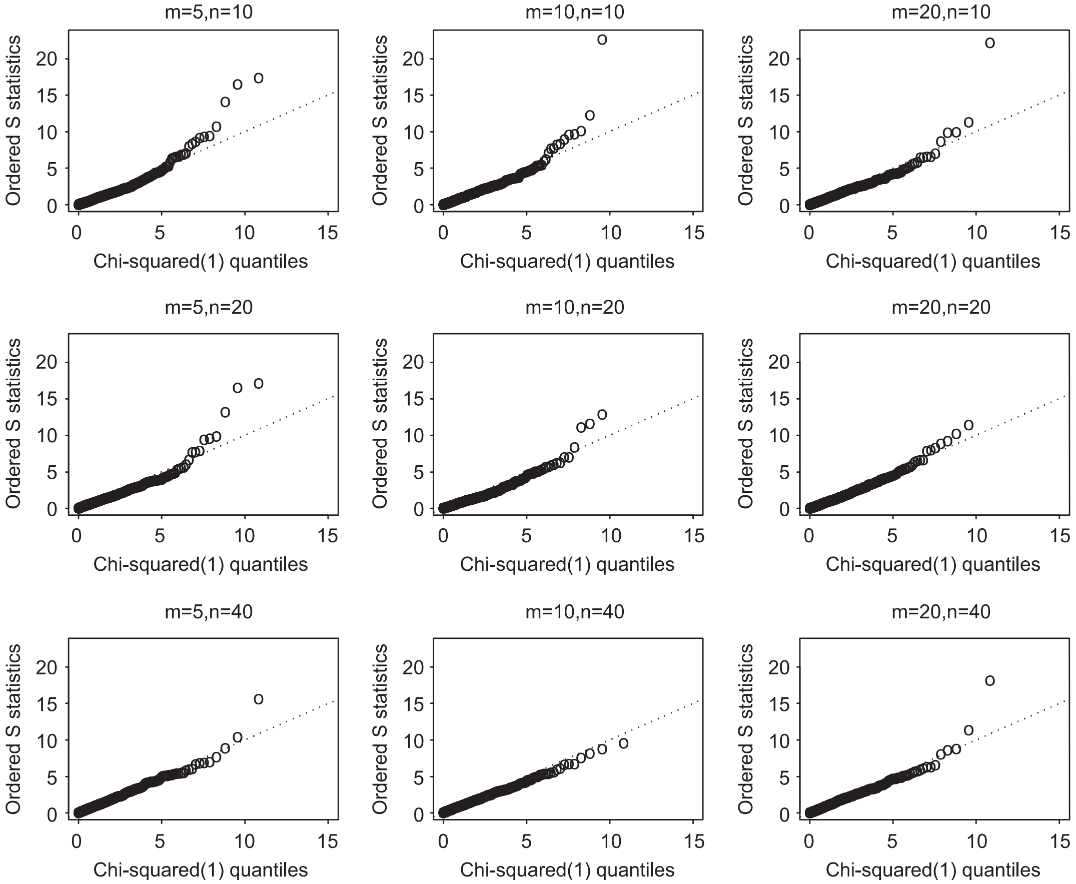

Figure 1. *Q–Q* plots of ordered score statistics against χ^2_1 quantiles based on 1000 replications generated from the Poisson mixed model (7) under $H_0 : \phi = 0$.

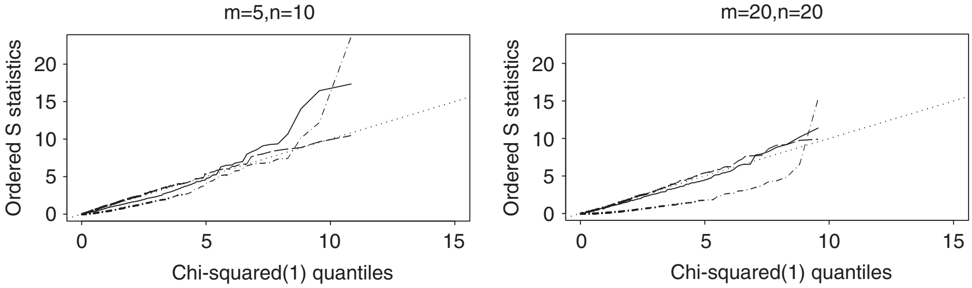

Figure 2. *Q–Q* plots of ordered score statistics against χ^2_1 quantiles based on 1000 replications generated from the Poisson mixed model (7) under $H_0 : \phi = 0$ with $\sigma_u = 0.1$ (broken line), 0.5 (solid line), 1.0 (broken–dotted line).

Table I. Empirical power of the score statistic S based on 1000 replications generated from a ZIP mixed model.

		$\phi = 0.25$			$\phi = 0.45$		
n	m	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
10	5	0.411	0.320	0.166	0.653	0.547	0.363
	10	0.587	0.492	0.317	0.885	0.830	0.680
	20	0.847	0.771	0.597	0.991	0.987	0.956
20	5	0.706	0.626	0.461	0.909	0.875	0.769
	10	0.899	0.835	0.704	0.995	0.985	0.965
	20	0.989	0.974	0.916	1.000	1.000	0.999
40	5	0.873	0.818	0.675	0.989	0.983	0.963
	10	0.991	0.983	0.950	1.000	1.000	0.998
	20	1.000	1.000	0.999	1.000	1.000	1.000

4.2. Power

Since zero-deflation is much less commonly encountered in practice, we investigate the power of the test for the zero-inflation situation only. Performance of the test procedure is evaluated under the ZIP mixed regression model, with linear predictor of the mean component having the same specification as (7) and associated parameter values as defined in Section 4.1. The empirical power of the test (for given ϕ) is calculated using the estimated upper tail probabilities of S at $\chi_1^2(1-\alpha)$ under an alternative hypothesis $H_1 : \phi \neq 0$, i.e.

$$P\{S > \chi_1^2(1-\alpha)\} \approx \sum_{k=1}^{1000} I[S_k > \chi_1^2(1-\alpha)|H_1]/1000$$

where S_k is the observed score statistic at the kth replicated trial, $k = 1, \ldots, 1000$. Both small ($\phi = 0.25$) and relatively large ($\phi = 0.45$) degrees of zero-inflation are considered, together with commonly adopted significance levels $\alpha = 0.10, 0.05, 0.01$.

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

5. RECURRENT UTI

UTI is one of the most common bacterial infections affecting women aged 60 years and above [17]. A retrospective cohort study was conducted in 2003 to determine the risk factors associated with recurrent UTI among elderly women in residential aged-care facilities. Eligibility criteria for the subjects were defined to be female residents aged 60 years or above with an institutionalization period of at least 6 months. A total of $N = 201$ subjects satisfying the selection criteria were recruited from $m = 6$ randomly selected aged-care institutions located in the Perth metropolitan area of Western Australia. The outcome variable was the number of

Table II. Frequency distribution of UTI counts of $N = 201$ residents by institution.

	UTI count											
Institution	0	1	2	3	4	5	6	7	8	9	10	>10
1	29	4	2	0	0	1	0	0	0	0	0	0
2	17	8	2	0	0	0	0	1	0	0	0	0
3	22	5	8	5	3	1	2	0	0	0	1	0
4	19	4	3	1	1	2	1	1	0	0	1	1
5	11	5	2	1	1	1	1	1	0	0	0	0
6	10	5	6	3	4	0	1	2	2	0	0	0
Frequency	108	31	23	10	9	5	5	5	2	0	2	1

UTI episodes during the 2 years follow-up period. Available covariates are binary variables indicating the presence $(=1)$ or absence $(=0)$ of $x_1 =$ history of prior UTI, $x_2 =$ urinary incontinence, $x_3 =$ anatomical abnormalities, and $x_4 =$ immuno-compromised. These variables were chosen because they are either established or postulated risk factors for UTI. Both histories of prior UTI and urinary incontinence were identified by reviewing the residents' past medical history. Subjects with anatomical abnormalities of the urinary tract included the presence of renal stones, strictures, cysts and obstruction. Information on these was documented from their doctor's referral letters. Women who were immuno-compromised received oral corticosteroids or chemotherapeutic agents and these were documented in their medication charts. Age and other co-morbidities such as diabetes mellitus, hysterectomy and prior stroke history were recorded, but these variables were excluded in subsequent analysis due to the high proportion of missing entries.

Table II gives the empirical frequency distribution of the UTI counts by institution, which ranged from zero to 17 episodes. For this cohort of 201 elderly women, 53.7 per cent experienced no episode of UTI, 24 per cent had a history of prior UTI, 39 per cent suffered from urinary incontinence, 8 per cent had anatomical abnormalities of the urinary tract, but only 3 per cent were immuno-compromised. Moreover, it is anticipated that women residing in the same institution were correlated in terms of contracting UTI because of their exposure to the same environment.

Parameter estimates from fitting the Poisson mixed model and the ZIP mixed model to the clustered data are presented in Table III(a) and (b). In addition to the four covariates, the duration of follow-up was included as an offset term in both models to adjust for the individual exposure. The results suggest that a history of UTI, presence of urinary incontinence and anatomical abnormalities, are all positively associated with the incidence of recurrent UTI, whereas the effect of being immuno-compromised is not significant. Based on the ZIP mixed model with constant zero-inflation parameter, the proportion of extra zeros is estimated to be 48 per cent. Meanwhile, the variation due to random cluster effects has reduced substantially after adjusting for the excess zeros in the ZIP mixed model. The score test statistic $S = 6.484$ is clearly significant (p-value $= 0.01$) with respect to the asymptotic χ_1^2 distribution, providing strong evidence of zero-inflation in this set of correlated count data. Furthermore, a likelihood ratio test of the Poisson mixed model against the alternative ZIP mixed model produces an approximate test statistic of 50.4 (p-value < 0.001) and agrees with the score test. Nevertheless, the score test is more appealing as it does not require a fit of the alternative ZIP mixed model.

Table III. Results from fitting Poisson mixed regression and ZIP mixed regression models to the recurrent UTI data.

			(c) Full ZIP mixed	
Parameter	(a) Poisson mixed, estimate (SE)	(b) ZIP mixed with constant ϕ, estimate (SE)	Poisson part, estimate (SE)	Logistic part, estimate (SE)
β_0	-6.810^* (0.248)	-5.839^* (0.140)	-5.781^* (0.134)	0.411 (0.351)
β_1	1.127^* (0.125)	0.778^* (0.134)	0.734^* (0.132)	-0.947^* (0.409)
β_2	0.322^* (0.121)	0.247^* (0.127)	0.216^* (0.128)	−0.290 (0.359)
β_3	0.771^* (0.163)	0.402^* (0.166)	0.376^* (0.169)	−1.163 (0.721)
β_4	0.470 (0.246)	0.179 (0.251)	0.166 (0.259)	−0.467 (1.040)
ϕ		0.479 (0.087)		
σ_u	0.538 (0.212)	0.095 (0.057)	0.091 (0.056)	
σ_v				0.587 (0.339)

$^*p$-value <0.05.

We next assess the homogeneity assumption concerning the zero-inflation parameter by allowing $h(\phi_{ij})$ in (3) as a logistic function of the covariates and random effects. Results of fitting the full ZIP mixed regression model, given in Table III(c), are similar to those of the restricted ZIP mixed model, with the exception of an additional significant factor identified from the logistic part. Specifically, patients with a history of prior UTI are 2.58 times more at risk of a recurrent UTI according to the full ZIP mixed model.

6. DISCUSSION

This paper proposes a score statistic for testing zero-inflation in correlated count data. The diagnostic procedure is useful for testing whether the observed high frequency of zeros renders unnecessary the fit of a more complex ZIP mixed regression model, and may be regarded as a generalization of the score test for zero-inflation in the standard Poisson situation [9]. The advantage of the score statistic lies in its computational convenience [12]; only a fit of the null Poisson mixed model is required and inference can be based on its asymptotic χ^2 distribution under the null hypothesis. The simulation results show that the test statistic performs well under a wide range of conditions. The example on recurrent UTI further demonstrates the practical applicability of the test procedure.

It should be remarked that the one-sided version of the test might be desirable for testing zero-inflation alone. If the alternative $\phi \neq 0$ is replaced by $\phi > 0$, according to Reference [18], the score test statistic should be modified as

$$S^* = U'\tilde{\mathfrak{I}}^{-1}U - \inf\{(U-b)'\tilde{\mathfrak{I}}^{-1}(U-b); b=(0,\ldots,0,b_0), b_0>0\}$$

where $U = U(\tilde{\beta}, \tilde{u}, \tilde{\tau}, 0)$ is given in Section 3. Specifically, the reference distribution of the score test statistic S^* is then asymptotically $0.5(\chi_0^2 + \chi_1^2)$ with p-value given by $0.5p\{\chi_1^2 > S^*\}$, i.e. the limiting distribution follows a mixture of a degenerate point mass at zero and a χ_1^2 distribution in equal mixing proportions.

The score test statistic is derived with respect to the underlying Poisson assumption. Depending on the nature of the response, alternative discrete probability distributions such as binomial may be adopted in (2) for the assessment of zero-inflation in other settings. In the presence of simultaneous zero-inflation and overdispersion, Ridout *et al.* [11] have developed a score test for testing ZIP against ZINB alternatives. It is worthwhile to extend their score test to the random effects setting for assessing over-dispersed correlated count data with extra zeros [19], findings of which will be reported elsewhere. Finally, multilevel ZIP [20] or multilevel ZINB regression models can be developed to analyse count data exhibiting a complex correlation structure due to multilevel clustering, once zero-inflation is confirmed by such score tests.

APPENDIX A

From the second derivatives of l evaluated at $\gamma=0$, entries of the expected Fisher information matrix $\Im(\beta,u,\tau,\gamma)$ under the null hypothesis H_0^* are obtained as follows:

$$\Im_{\beta\beta}=E\left[-\frac{\partial^2 l}{\partial\beta\partial\beta'}\right]=\sum_{i,j}\frac{\partial\eta_{ij}}{\partial\beta'}E\left[-\frac{\partial^2 l_{1ij}}{\partial\eta_{ij}^2}\right]\frac{\partial\eta_{ij}}{\partial\beta'}=\sum_{i,j}\exp(\eta_{ij})\frac{\partial\eta_{ij}}{\partial\beta'}\frac{\partial\eta_{ij}}{\partial\beta}=-T'BT$$

$$\Im_{uu}=E\left[-\frac{\partial^2 l}{\partial u\partial u'}\right]=\sum_{i,j}\frac{\partial\eta_{ij}}{\partial u'}E\left[-\frac{\partial^2 l_{1ij}}{\partial\eta_{ij}^2}\right]\frac{\partial\eta_{ij}}{\partial u}+\tau I_p=-P'BP+\tau I_p$$

$$\Im_{\beta u}=E\left[-\frac{\partial^2 l}{\partial u\partial\beta'}\right]=\sum_{i,j}\frac{\partial\eta_{ij}}{\partial\beta'}E\left[-\frac{\partial^2 l_{1ij}}{\partial\eta_{ij}^2}\right]\frac{\partial\eta_{ij}}{\partial u}=-T'BP$$

$$\Im_{\tau\tau}=E\left[-\frac{\partial^2 l}{\partial(\tau)^2}\right]=-\frac{m}{2}\tau^{-2}+\tau^{-3}\sum_i u_i^2$$

$$\Im_{\tau\gamma}=E\left[-\frac{\partial^2 l}{\partial\gamma\partial\tau}\right]=0$$

$$\Im_{\beta\tau}=E\left[-\frac{\partial^2 l}{\partial\beta\partial\tau}\right]=0$$

$$\Im_{u\tau}=E\left[-\frac{\partial^2 l}{\partial u\partial\tau}\right]=-\tau^{-2}u$$

$$\Im_{\gamma\gamma}=E\left[-\frac{\partial^2 l}{\partial\gamma^2}\right]=\sum_{i,j}\left[-1+\frac{E(I_{(y_{ij}=0)})}{f^2(0;\eta_{ij})}\right]=\sum_{i,j}\left[\frac{1}{f(0;\eta_{ij})}-1\right]$$

$$\Im_{\beta\gamma} = E\left[-\frac{\partial^2 l}{\partial\beta'\partial\gamma}\right] = \sum_{i,j}\frac{\partial\eta_{ij}}{\partial\beta'}E\left[-I_{(y_{ij}=0)}\frac{f(0;\eta_{ij})(-\exp(\eta_{ij}))}{[\gamma+f(0;\eta_{ij})]^2}\right]$$

$$= \sum_{i,j}\frac{\partial\eta_{ij}}{\partial\beta'}\frac{-\exp(\eta_{ij})}{f(0;\eta_{ij})}E(I_{(y_{ij}=0)})$$

$$= -\sum_{i,j}\exp(\eta_{ij})\frac{\partial\eta_{ij}}{\partial\beta'} = -T'B1_N$$

$$\Im_{u\gamma} = E\left[\frac{\partial}{\partial u}\left(-\frac{\partial l}{\partial\gamma}\right)\right] = E\left[\frac{\partial}{\partial u}\sum_{i,j}\left\{\frac{1}{1+\gamma} - I_{(y_{ij}=0)}\frac{1}{\gamma+f(0;\eta_{ij})}\right\}\right]$$

$$= E\left[\sum_{i,j}I_{(y_{ij}=0)}\frac{-f(0;\eta_{ij})\exp(\eta_{ij})}{(\gamma+f(0;\eta_{ij}))^2}\frac{\partial\eta_{ij}}{\partial u}\right] = -\sum_{i,j}\exp(\eta_{ij})\frac{\partial\eta_{ij}}{\partial u} = -P'B1_N$$

with $N\times q$ matrix

$$T = \left(\frac{\partial\eta_{11}}{\partial\beta},\ldots,\frac{\partial\eta_{1n_1}}{\partial\beta},\ldots,\frac{\partial\eta_{m1}}{\partial\beta},\ldots,\frac{\partial\eta_{mn_m}}{\partial\beta}\right)'$$

$N\times m$ matrix

$$P = \left(\frac{\partial\eta_{11}}{\partial u},\ldots,\frac{\partial\eta_{1n_1}}{\partial u},\ldots,\frac{\partial\eta_{m1}}{\partial u},\ldots,\frac{\partial\eta_{mn_m}}{\partial u}\right)'$$

and $N\times N$ matrices $B = \text{diag}\{-\exp(\eta_{ij})\}$. Here 1_N denotes an $N\times 1$ vector of ones.

ACKNOWLEDGEMENTS

The authors are grateful to Jayne Wee, School of Pharmacy, Curtin University of Technology, for providing the urinary tract infection data. Thanks are also due to two anonymous reviewers for helpful comments and suggestions. This research was supported by grants from the National Health & Medical Research Council of Australia, the Australian Research Council, and the Research Grants Council of Hong Kong.

REFERENCES

1. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**:1–14.
2. Böhning D. Zero-inflated Poisson models and C.A. Man: a tutorial collection of evidence. *Biometrical Journal* 1998; **40**:833–843.
3. Ridout M, Demétrio CGB, Hinde J. Models for count data with many zeros. *Proceedings of the XIXth International Biometrics Conference*, Cape Town, 1998; 179–192.
4. 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.
5. Lee AH, Wang K, Yau KKW. Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal* 2001; **43**:963–975.
6. 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.

7. 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.
8. McGilchrist CA. Estimation in generalized mixed models. *Journal of the Royal Statistical Society*, *Series B* 1994; **56**:61–69.
9. Van den Broek J. A score test for zero-inflation in a Poisson distribution. *Biometrics* 1995; **51**:738–743.
10. Deng D, Paul SR. Score tests for zero inflation in generalized linear models. *The Canadian Journal of Statistics* 2000; **27**:563–570.
11. 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**:219–223.
12. Jansakul N, Hinde JP. Score tests for zero-inflated Poisson models. *Computational Statistics and Data Analysis* 2002; **40**:75–96.
13. Dietz K, Böhning D. On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis* 2000; **34**:441–459.
14. McLachlan GJ. On the EM algorithm for overdispersed count data. *Statistical Methods in Medical Research* 1997; **6**:76–98.
15. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions.* Wiley: New York, 1997.
16. Cox DR, Hinkley DV. *Theoretical Statistics.* Chapman & Hall: New York, 1979.
17. Nicolle LE. Urinary tract infection in geriatric and institutionalised patients. *Current Opinion in Urology* 2002; **12**:51–55.
18. Silvapulle MJ, Silvapulle P. A score test against one-sided alternatives. *Journal of the American Statistical Association* 1995; **90**:342–349.
19. 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.
20. 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* 2005, in press.