The Canadian Journal of Statistics
Vol. 30, No. 3, 2002, Pages 1–16
La revue canadienne de statistique

Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models

Daniel B. HALL and Kenneth S. BERENHAUT

Key words and phrases: Beta-binomial distribution; excess zeros; mixed effects; negative binomial distribution; random effects; repeated measures; variance components.

MSC 2000: Primary 62J12; secondary 62F03.

Abstract: Hall (2000) has described zero-inflated Poisson and binomial regression models that include random effects to account for excess zeros and additional sources of heterogeneity in the data. The authors of the present paper propose a general score test for the null hypothesis that variance components associated with these random effects are zero. For a zero-inflated Poisson model with random intercept, the new test reduces to an alternative to the overdispersion test of Ridout, Demétrio & Hinde (2001). The authors also examine their general test in the special case of the zero-inflated binomial model with random intercept and propose an overdispersion test in that context which is based on a beta-binomial alternative.

Tests scores d'hétérogénéité et de surdispersion dans des modèles de régression binomiaux et de Poisson avec surplus de zéros

Résumé : Hall (2000) a décrit des modèles de régression binomiaux et de Poisson dans lesquels des effets aléatoires servent à expliquer certaines sources d'hétérogénéité dans les données, dont un surplus de zéros. Les auteurs du présent article proposent un test score général permettant de vérifier si les composantes de la variance associées à ces effets aléatoires sont nulles. Pour un modèle de Poisson à surplus de zéros et à ordonnée aléatoire, le nouveau test se compare au test de surdispersion de Ridout, Demétrio & Hinde (2001). Les auteurs étudient en outre leur test général dans le cadre du modèle binomial à surplus de zéros et à ordonnée aléatoire, pour lequel ils proposent un test de surdispersion adapté à des contre-hypothèses de type bêta-binomial.

1. INTRODUCTION

In many data sets involving counts, large frequencies of zeros are observed relative to what is predicted by models based on standard distributional assumptions. In the case of bounded counts, when the response can be viewed as the number of successes out of a fixed number of trials, the standard distribution for regression modelling is the binomial. In the case of unbounded counts, Poisson regression models are standard. Recently, it has become popular to model such data using regression models based on an assumption that the response is generated by a mixture of one of the standard count distributions with a degenerate distribution with point mass of 1 at zero. These zero-inflated binomial (ZIB) and zero-inflated Poisson (ZIP) models have a history that goes back at least to Mullahy (1986), who discussed the ZIP model. Ridout, Demétrio & Hinde (1998) provide a review of the literature on ZIP and other closely related zero-inflation models for unbounded counts through 1998. More recently, Hall (2000) and Vieira, Hinde & Demétrio (2000) have introduced ZIB models.

Most of this work has treated the regression parameters in these models as fixed unknown parameters and has involved an assumption of independence among all of the observations in the data set. Recently Hall (2000) has considered the clustered-data case and shown how mixed-effects versions of ZIP and ZIB models appropriate to this case can be specified and fit. In this paper, we consider the problem of testing for the necessity of random effects in a ZIP or ZIB model. That is, we test the null hypothesis that the variance and covariance components associated with random effects in a ZIP-mixed or ZIB-mixed model are equal to zero. The test we propose is a score test similar to that of Lin (1997), whose test was presented in the context of generalized linear mixed-effects models. In that context, Lin's test reduces to certain tests of overdispersion (Cox 1983; Dean 1992) in special cases. Similarly, our test reduces to a test of overdispersion relative to a ZIP or ZIB model in a non-clustered data context. In the ZIP case, this overdispersion test provides an alternative to the score test of Ridout, Demétrio & Hinde (2001) for misspecification of a ZIP model relative to a zero-inflated negative binomial (ZINB) alternative. In the ZIB case, we present two overdispersion tests, one that corresponds to a special case of our general test and a second that is based on a zero-inflated beta-binomial (ZIBB) alternative. Score tests for zero-inflation in Poisson and binomial regression models have been investigated by van den Broek (1995) and Deng & Paul (2000).

2. ZERO-INFLATED MIXED MODELS

2.1. The ZIP case.

Suppose we observe an $N-$dimensional response vector $\mathbf{y}$ containing data from K independent clusters, so that $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_K')'$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{it_i})'$. We assume that, conditional on a $q-$dimensional vector of random effects $\boldsymbol{\beta}_i$, the random variable Y_{ij} associated with the observation y_{ij} follows the distribution

$$Y_{ij}|\boldsymbol{\beta}_i \sim \begin{cases} 0, & \text{with probability } p_{ij}; \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } 1 - p_{ij}. \end{cases}$$

In addition, we model $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{it_i})'$ and $\mathbf{p}_i = (p_{i1}, \ldots, p_{it_i})'$ as functions of covariates and regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ through a log-linear generalized linear mixed model and a logistic generalized linear model, respectively, as follows

$$\begin{aligned} \log(\boldsymbol{\lambda}_i) &= \mathbf{B}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \\ \text{logit}(\mathbf{p}_i) &= \mathbf{G}_i\boldsymbol{\gamma}, \end{aligned}$$

for $i = 1, \ldots, K$. Here, $\mathbf{B} = (\mathbf{B}_1', \ldots, \mathbf{B}_K')'$, $\mathbf{W} = (\mathbf{W}_1', \ldots, \mathbf{W}_K')'$, and $\mathbf{G} = (\mathbf{G}_1', \ldots, \mathbf{G}_K')'$ are design matrices of dimension $N \times p$, $N \times q$ and $N \times r$, respectively, and we assume $\mathbf{b}_1, \ldots, \mathbf{b}_K$ are independent and identically distributed (i.i.d.) $q-$variate random vectors following distribution F with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{D}$.

It is convenient to utilize an unconstrained parameterization for $\mathbf{D}$. In particular, using the Cholesky parameterization (Pinheiro & Bates 1996), we can rewrite the log-linear portion of the model as

$$\log(\boldsymbol{\lambda}_i) = \mathbf{B}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{D}^{T/2}(\boldsymbol{\theta})\mathbf{b}_i^*,$$

where $\mathbf{b}_1^*, \ldots, \mathbf{b}_K^*$ are i.i.d. with mean $\mathbf{0}$ and identity variance-covariance matrix, and $\mathbf{D}^{T/2}(\boldsymbol{\theta})$ is the transpose of $\mathbf{D}^{1/2}(\boldsymbol{\theta})$, the upper triangular Cholesky factor of $\mathbf{D}$. We parameterize the variance-covariance matrix of $\mathbf{b}_i$ through an unconstrained vector-valued parameter $\boldsymbol{\theta}$ that contains the nonzero elements of $\mathbf{D}^{1/2}$. In this paper, we are interested in tests of the hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}$. The unconstrained parameterization is not necessary for the results of this paper, but we do assume the parametrization is such that $\mathbf{D}(\boldsymbol{\theta}) = \mathbf{0}$ if $\boldsymbol{\theta} = \mathbf{0}$.

As an example of such a model, suppose that we observe a count-valued response with many zeros from a two-period, two-treatment crossover design. Such data would be clustered at the subject level. We may be interested in exploring the possibility that there is baseline heterogeneity from subject to subject in addition to heterogeneity in the treatment effects across subjects. If so, it would be natural to entertain a ZIP model with linear predictor $\boldsymbol{\eta}_i = \log(\boldsymbol{\lambda}_i)$ containing a random intercept and random treatment effect which could be allowed to covary. That is, $\boldsymbol{\eta}_i$

would be of the form $\mathbf{B}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{D}^{T/2}(\boldsymbol{\theta})\mathbf{b}_i^*$, where $\mathbf{W}_i$ is $t_i \times 2$ containing a column of ones and $\mathbf{trt}_i$, a treatment indicator, and

$$\mathbf{b}_i^* = \begin{pmatrix} b_{i1}^* \\ \\ b_{i2}^* \end{pmatrix}, \quad \mathbf{D}^{T/2} = \begin{pmatrix} \theta_1 & 0 \\ \\ \theta_2 & \theta_3 \end{pmatrix}$$

so that $\boldsymbol{\eta}_{ij}$ has variance $\theta_1^2 + (2\theta_1\theta_2 + \theta_2^2 + \theta_3^2)\mathrm{trt}_{ij}$.

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{\theta}')'$ be the combined parameter vector. The log-likelihood for the ZIP-mixed model is

$$\ell(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i=1}^{K} \log \int \left\{ \prod_{j=1}^{t_i} \mathrm{P}(Y_{ij} = y_{ij}|\mathbf{b}_i) \right\} dF(\mathbf{b}_i), \tag{1}$$

where the integral is $q-$dimensional over $(-\infty, \infty) \times \cdots \times (-\infty, \infty)$ (q times), and $\mathrm{P}(Y_{ij} = y_{ij}|\mathbf{b}_i)$ is given by

$$\left\{p_{ij} + (1-p_{ij})e^{-\lambda_{ij}}\right\}^{I_{\{y_{ij}=0\}}} \left\{(1-p_{ij})e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}/(y_{ij}!)\right\}^{1-I_{\{y_{ij}=0\}}}.$$

Here, $I_{\{A\}} = 1$ if condition A holds, $I_{\{A\}} = 0$ otherwise.

Hall (2000) has described how this log-likelihood can be maximized using the EM algorithm combined with Gaussian quadrature to fit the model in the case $q = 1$, where the random effect is assumed to be normally distributed. Extension to vector-valued $\mathbf{b}_i$ is straightforward, proceeding along the lines described in Pinheiro & Bates (1995). However, for q greater than 2 or 3, the computations become impractical. An example of a model fit with this approach with $q = 2$ appears in van Iersel, Oetting, Hall & Kang (2001). An improvement on the approach of Hall (2000) would be to use adaptive Gaussian quadrature, as described by Pinheiro & Bates (1995), in the EM algorithm. In principle, SAS' PROC NLMIXED (Wolfinger 1999) can be used to fit this model without the EM algorithm using any one of several optimization algorithms, including the Newton-Raphson method and various quasi-Newton methods combined with Gaussian quadrature, adaptive Gaussian quadrature or one of the other integration approximation methods described by Pinheiro & Bates (1995). However, even with the convenience of this package, computational problems leading to non-convergence can be difficult obstacles to overcome. A natural question to address, therefore, is whether or not the data support the inclusion of random effects. If not, then the somewhat cumbersome methodology of the mixed-effects case can be avoided.

2.2. The ZIB case.

In the ZIB case, the observation y_{ij} is a bounded count which can be thought of as the number of successes occurring out of n_{ij} trials. In the ZIB-mixed model, it is assumed that

$$Y_{ij}|\mathbf{b}_i \sim \begin{cases} 0, & \text{with probability } p_{ij}; \\ \\ \text{Binomial}(n_{ij}, \pi_{ij}), & \text{with probability } 1 - p_{ij}. \end{cases}$$

In addition, the mixing probabilities $\mathbf{p}_i$ and success probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{it_i})'$ are assumed to follow logistic models:

$$\begin{aligned} \mathrm{logit}(\boldsymbol{\pi}_i) &= \mathbf{B}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i, \\ \mathrm{logit}(\mathbf{p}_i) &= \mathbf{G}_i\boldsymbol{\gamma}, \end{aligned}$$

for $i = 1, \ldots, K$. The log-likelihood for this model is as given in (1), but with

$$\begin{aligned} \mathrm{P}(Y_{ij} = y_{ij}|\mathbf{b}_i) &= \{p_{ij} + (1-p_{ij})(1-\pi_{ij})^{n_{ij}}\}^{I_{\{y_{ij}=0\}}} \\ &\quad \times \left\{(1-p_{ij})\binom{n_{ij}}{y_{ij}}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{n_{ij}-y_{ij}}\right\}^{1-I_{\{y_{ij}=0\}}}. \end{aligned}$$

3. A GENERAL SCORE TEST FOR HETEROGENEITY

In either the ZIP case or the ZIB case, let $\boldsymbol{\delta} = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$ and define $\ell_{ij}(\boldsymbol{\delta}; \mathbf{b}_i) = \log\{\mathrm{P}(Y_{ij} = y_{ij}|\mathbf{b}_i)\}$. Then, following Lin (1997), a Laplace-like approximation to $\ell(\boldsymbol{\psi}; \mathbf{y})$ can be obtained by taking a two-term Taylor expansion of the integrand in (1) about $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$. The resulting expression for the log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\psi}; \mathbf{y}) &= \sum_{i=1}^{K}\sum_{j=1}^{t_i} \ell_{ij}(\boldsymbol{\delta}; \mathbf{0}) \\ &+\frac{1}{2}\sum_{i=1}^{K} \mathrm{tr}\left[\left\{\left(\frac{\partial \ell_i(\boldsymbol{\delta}; \mathbf{0})}{\partial \boldsymbol{\eta}_i}\right)\left(\frac{\partial \ell_i(\boldsymbol{\delta}; \mathbf{0})}{\partial \boldsymbol{\eta}_i'}\right) + \frac{\partial^2 \ell_i(\boldsymbol{\delta}; \mathbf{0})}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i'}\right\} \mathbf{W}_i \mathbf{D}(\boldsymbol{\theta}) \mathbf{W}_i'\right] + \mathrm{o}(||\boldsymbol{\theta}||).\end{aligned}$$

Here $\ell_i(\boldsymbol{\delta}; \mathbf{b}_i)$ is the contribution to the log-likelihood from the ith cluster, $\partial \ell_i(\boldsymbol{\delta}; \mathbf{0})/\partial \boldsymbol{\eta}_i$ is the $t_i \times 1$ vector with jth element $\partial \ell_{ij}(\boldsymbol{\delta}; \mathbf{0})/\partial \eta_{ij}$, and $\partial^2 \ell_i(\boldsymbol{\delta}; \mathbf{0})/(\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i)$ is a $t_i \times t_i$ diagonal matrix with jth diagonal element $\partial^2 \ell_{ij}(\boldsymbol{\delta}; \mathbf{0})/(\partial \eta_{ij}^2)$.

The score test statistic for $\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}$ is given by

$$T^2 = \mathbf{U}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}}_0)' \mathbf{I}^{\boldsymbol{\theta\theta}}(\hat{\boldsymbol{\delta}}_0) \mathbf{U}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}}_0), \qquad (2)$$

where $\hat{\boldsymbol{\delta}}_0$ is the maximum likelihood estimate of $\boldsymbol{\delta}$ under $\mathcal{H}_0$, $\mathbf{U}_{\boldsymbol{\theta}}$ is the efficient score vector for $\boldsymbol{\theta}$, and $\mathbf{I}^{\boldsymbol{\theta\theta}}$ is the submatrix of the inverse Fisher information matrix corresponding to $\boldsymbol{\theta}$. That is, if we partition the Fisher information as

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{\boldsymbol{\delta\delta}} & \mathbf{I}_{\boldsymbol{\delta\theta}} \\ \mathbf{I}_{\boldsymbol{\delta\theta}}' & \mathbf{I}_{\boldsymbol{\theta\theta}} \end{pmatrix},$$

then $\mathbf{I}^{\boldsymbol{\theta\theta}} = (\mathbf{I}_{\boldsymbol{\theta\theta}} - \mathbf{I}_{\boldsymbol{\delta\theta}}' \mathbf{I}_{\boldsymbol{\delta\delta}}^{-1} \mathbf{I}_{\boldsymbol{\delta\theta}})^{-1}$. Both $\mathbf{U}_{\boldsymbol{\theta}}$ and $\mathbf{I}^{\boldsymbol{\theta\theta}}$ are evaluated at $\boldsymbol{\theta} = \mathbf{0}$ and $\boldsymbol{\delta} = bd\hat{e}lta_0$. Under the regularity conditions of Proposition 1 of Lin (1997), T^2 will have an asymptotic chi-square distribution on $\dim(\boldsymbol{\theta})$ degrees of freedom under the null hypothesis that $\boldsymbol{\theta}$ is equal to $\mathbf{0}$.

Note that normality of the random effects' distribution is not necessary for our score test. We assume only that the first two moments of $\mathbf{b}_i$ are $\mathbf{0}$ and $\mathbf{D}(\boldsymbol{\theta})$, respectively, and that higher order moments are of order $o(||\boldsymbol{\theta}||)$. As noted by Lin (1997), this condition holds for $\mathbf{b}_i$ drawn from the exponential family or from a mixture of exponential family distributions.

In general, the score vector $\mathbf{U}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}}_0)$ is $\dim(\boldsymbol{\theta}) \times 1$ and has jth element

$$U_{\boldsymbol{\theta}_j}(\hat{\boldsymbol{\delta}}_0) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{tr}\left[\left\{\left(\frac{\partial \ell_i(\boldsymbol{\delta}; \mathbf{0})}{\partial \boldsymbol{\eta}_i}\right)\left(\frac{\partial \ell_i(\boldsymbol{\delta}; \mathbf{0})}{\partial \boldsymbol{\eta}_i'}\right) + \frac{\partial^2 \ell_i(\boldsymbol{\delta}; \mathbf{0})}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i'}\right\} \mathbf{W}_i \dot{\mathbf{D}}_j \mathbf{W}_i'\right]\Bigg|_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}_0}, \qquad (3)$$

where $\dot{\mathbf{D}}_j = \partial \mathbf{D}/\partial \boldsymbol{\theta}_j\big|_{\boldsymbol{\theta}=0}$. In both the ZIP case and the ZIB case, $\ell_{ij}(\boldsymbol{\delta}; \mathbf{0})$ takes the form

$$I_{\{y_{ij}=0\}} \log(\rho_{0,ij}) + (1 - I_{\{y_{ij}=0\}}) \log\{(1 - p_{ij}) g(y_{ij}; \eta_{ij})\},$$

where $\rho_{0,ij} = p_{ij} + (1 - p_{ij}) g(0; \eta_{ij})$ equals $\mathrm{P}(Y_{ij} = 0)$ under $\mathcal{H}_0$; and $g(y_{ij}; \eta_{ij}) = c(y_{ij}) \exp\{y_{ij}\eta_{ij} - b(\eta_{ij})\}$ is a probability function of exponential family form with canonical link, so that the canonical parameter is equal to the linear predictor η_{ij}. The quantity $\partial \ell_i(\boldsymbol{\delta}; \mathbf{0})/\partial \boldsymbol{\eta}_i$ in (3) has jth element equal to $\{y_{ij} - b'(\eta_{ij})\} + I_{\{y_{ij}=0\}} b'(\eta_{ij}) p_{ij}/\rho_{0,ij}$. In addition, $\partial^2 \ell_i(\boldsymbol{\delta}; \mathbf{0})/(\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i')$ is a diagonal matrix with jth diagonal element equal to

$$-b''(\eta_{ij}) + I_{\{y_{ij}=0\}} p_{ij} [b''(\eta_{ij}) + \{b'(\eta_{ij})\}^2 (1 - p_{ij}/\rho_{0,ij})]/\rho_{0,ij}.$$

The general form of the information matrix is notationally complex and is given in the Appendix.

Note that $\eta_{ij} = \log(\lambda_{ij})$ and $b(\eta_{ij}) = e^{\eta_{ij}}$ in the ZIP case; and $\eta_{ij} = \operatorname{logit}(\pi_{ij})$, and $b(\eta_{ij}) = n_{ij}\log(1+e^{\eta_{ij}})$ in the ZIB case. Substituting these expressions into (3) and into the expressions in the Appendix for the information matrix, we obtain the general score test for these models from (2). Special cases of this test corresponding to tests of overdispersion in the unclustered cases (i.e., cluster sizes of $t_i = 1$ for all i) of the ZIP and ZIB models are examined in more detail in the following section.

In ordinary (not zero-inflated) regression contexts, other authors have noted the poor quality of the asymptotic chi-square approximation to the distribution of similarly derived score tests for small samples. Dean & Lawless (1989), Dean (1992), and Lin (1997) have all proposed small-sample adjustments for their tests of homogeneity and overdispersion, which vastly improve the chi-square approximation. We propose a similar adjustment in the zero-inflated context for T, which generalizes the small-sample adjusted tests of Lin (1997) and Dean (1992). The adjustment is based on the approximation

$$\mathrm{E}[\{(y_{ij} - b'(\hat{\eta}_{ij})\}^2] \approx p_{ij}\left[\{b'(\eta_{ij})\}^2 - b''(\eta_{ij})\right] + (1-h_{ij})b''(\eta_{ij}),$$

where h_{ij} is the (i,j)th element of the matrix

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{B}(\mathbf{B}'\mathbf{W}\mathbf{B})^{-1}\mathbf{B}'\mathbf{W}^{1/2},$$

where $\mathbf{W}$ is a diagonal matrix with (i,j)th diagonal element $(1-z_{ij})b''(\eta_{ij})$ and $z_{ij} = \{1 + e^{-\mathbf{G}_{ij}\boldsymbol{\gamma}}g(0;\eta_{ij})\}^{-1}$. The small-sample adjusted test takes the form

$$(T^c)^2 = \mathbf{U}^c_{\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}}_0)'\mathbf{I}^{\boldsymbol{\theta}\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}}_0)\mathbf{U}^c_{\boldsymbol{\theta}}(\hat{\boldsymbol{\delta}}_0),$$

where $\mathbf{U}^c_\theta(\hat{\delta}_0)$ is as in (3), but with $\partial^2\ell_i(\delta;\mathbf{0})/(\partial\boldsymbol{\eta}_i\partial\boldsymbol{\eta}_i')$ replaced by a diagonal matrix with jth diagonal element equal to

$$-(1-h_{ij})b''(\eta_{ij}) + I_{\{y_{ij}=0\}}p_{ij}[b''(\eta_{ij}) + \{b'(\eta_{ij})\}^2(1-p_{ij}/\rho_{0,ij})]/\rho_{0,ij}.$$

We investigate the power and size properties of the adjusted and unadjusted test via simulation in Section 6.

4. SPECIAL CASES: TESTING FOR OVERDISPERSION

Throughout this section, we assume $t_i = 1$ for all i, so that $K = N$.

4.1. The ZIP case.

Recently Ridout, Demétrio & Hinde (2001) have presented a score test for testing a ZIP model against a ZINB alternative. These authors point out the more serious consequences (inconsistent parameter estimators) resulting from misspecification of the nondegenerate portion of the mixture distribution in a ZIP model than in a standard Poisson regression model. This result provides good motivation for a test of overdispersion such as the one they propose from the relatively restrictive Poisson form of a zero-inflated model for count data.

The ZINB alternative to the ZIP can be regarded as a ZIP-mixed model with $\log(\lambda_i) = \mathbf{B}_i\boldsymbol{\beta} + b_i$, where e^{b_i} follows a gamma distribution. The gamma distributional assumption on the random effect is mathematically convenient in that it is conjugate for the Poisson and thus leads to a closed form expression for the marginal distribution of y_i. However, this assumption does not easily generalize to vector-valued random effects. In applying our general score test to this context, we assume only that $b_1, \ldots, b_N$ are i.i.d. with mean 0, variance θ and higher moments that are $\mathrm{o}(\theta)$.

Under these assumptions, our score test from Section 3 applies and leads to a score statistic equal to

$$U_\theta(\hat{\boldsymbol{\delta}}_0) = \frac{1}{2}\sum_{i=1}^{N}\left\{\left(y_i - \hat{\lambda}_i\right)^2 - \hat{\lambda}_i - I_{\{y_i=0\}}\hat{\lambda}_i\left(\hat{\lambda}_i - 1\right)\hat{p}_i/\hat{\rho}_{0,i}\right\}. \qquad (4)$$

In this case $I^{\theta\theta}(\hat{\boldsymbol{\delta}}_0)$ is a scalar which can be obtained from the information matrix $\mathbf{I}(\boldsymbol{\delta})$ evaluated at $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}_0$. Typical elements of $\mathbf{I}(\boldsymbol{\delta})$ are

$$\begin{aligned}
I_{\theta\theta} &= \frac{1}{4}\sum_{i=1}^{N}\left\{(1-p_i)(2\lambda_i^2+\lambda_i) - p_i(\lambda_i^2-\lambda_i)^2\kappa_i\right\},\\
I_{\theta\gamma_j} &= \frac{1}{2}\sum_{i=1}^{N} p_i G_{ij}\lambda_i(\lambda_i-1)\kappa_i,\\
I_{\theta\beta_j} &= \frac{1}{2}\sum_{i=1}^{N} \lambda_i B_{ij}\left\{(1-p_i)+\lambda_i(\lambda_i-1)p_i\kappa_i\right\},\\
I_{\gamma_j\gamma_k} &= \sum_{i=1}^{N} p_i^2 G_{ij}G_{ik}(\rho_{0,i}^{-1}-1),\\
I_{\gamma_j\beta_k} &= -\sum_{i=1}^{N} p_i\lambda_i G_{ij}B_{ik}\kappa_i,\\
I_{\beta_j\beta_k} &= \sum_{i=1}^{N} \lambda_i B_{ij}B_{ik}\left\{(1-p_i)-p_i\lambda_i\kappa_i\right\},
\end{aligned}$$

where $\kappa_i = (1 - p_i/\rho_{0,i})$. The one-sided score test of $\mathcal{H}_0 : \theta = 0$ is then

$$T = U_\theta(\hat{\boldsymbol{\delta}}_0)\sqrt{I^{\theta\theta}(\hat{\boldsymbol{\delta}}_0)},$$

which has an asymptotic standard normal distribution under $\mathcal{H}_0$.

Ridout, Demétrio & Hinde (2001) provide some intuition for their score function as a test statistic for overdispersion by writing it as a linear combination of terms that compare first and second sample moments and the zero frequency with the corresponding theoretical values under the ZIP model. This is possible with our alternative score statistic (4) as well. Let $\mu_i = (1-p_i)\lambda_i$ and $\sigma_i^2 = \mu_i\{1 + p_i\mu_i/(1-p_i)\}$ denote the mean and variance of y_i under $\mathcal{H}_0$. Then $U_\theta(\hat{\boldsymbol{\delta}}_0)$ can be written as

$$\frac{1}{2}\sum_{i=1}^{N}\left[\left\{(y_i-\hat{\mu}_i)^2-\hat{\sigma}_i^2\right\} - \frac{2\hat{\mu}_i\hat{p}_i}{1-\hat{p}_i}(y_i-\hat{\mu}_i) + \frac{\hat{\mu}_i\hat{p}_i}{(1-\hat{p}_i)\hat{\rho}_{0,i}}\left(1-\frac{\hat{\mu}_i}{1-\hat{p}_i}\right)(I_{\{y_i=0\}}-\hat{\rho}_{0,i})\right],$$

where hats indicate evaluation at the MLE $\hat{\boldsymbol{\delta}}_0$ under the null hypothesis.

In the context of overdispersion testing in Poisson regression models, Dean (1992) points out the similarity between a log-linear mixed effects alternative and a multiplicative random effects, or negative binomial, alternative. These two alternative hypotheses lead to two overdispersion tests, P_A and P_B in the notation of Dean's paper, that are quite similar. In fact they coincide for log-linear Poisson regression models in which the linear predictor includes an intercept. The test of Ridout, Demétrio & Hinde (2001) and our test reduce to P_A and P_B, respectively, in the case of no zero-inflation and similar comments and results apply in our ZIP context. For the parameterization of the negative binomial corresponding to $c = 1$ in Ridout, Demétrio & Hinde (2001), the difference between our score function and theirs is

$$\frac{1}{2}\sum_{i=1}^{N}\left(y_i - \hat{\lambda}_i + I_{\{y_i=0\}}\hat{\lambda}_i\hat{p}_i/\hat{\rho}_{0,i}\right).$$

This difference will be zero and the two tests will coincide in a ZIP model with log link for $\boldsymbol{\lambda}$ that includes an intercept in $\mathbf{B}\boldsymbol{\beta}$.

4.2. The ZIB case.

Ridout, Demétrio & J. Hinde (2001) consider only overdispersion in the ZIP model in their paper. However, their approach for deriving a score test for overdispersion can easily be applied to the ZIB model by considering a ZIBB alternative. In this section, we present a score test based on this alternative, and we also apply our general score test of Section 3 to the same setting.

The ZIBB alternative we consider assumes

$$\mathrm{P}(Y_i = y_i) = \begin{cases} p_i + (1-p_i)g(0;\theta), & \text{if} y_i = 0; \\ (1-p_i)g(y_i;\theta), & \text{if} y_i > 0, \end{cases}$$

where

$$g(y_i;\theta) = \binom{n_i}{y_i} \prod_{j=0}^{y_i-1}(\pi_i + \theta j) \prod_{j=0}^{n_i-y_i-1}(1-\pi_i+\theta j) \Big/ \prod_{j=0}^{n_i-1}(1+\theta j).$$

Here, $g(y_i;\theta)$ is the beta-binomial probability function with overdispersion parameter θ (see Prentice 1986). Again, we are interested in testing $\mathcal{H}_0 : \theta = 0$, under which the ZIBB model reduces to the ZIB model.

The score statistic for testing $\mathcal{H}_0$ in the ZIBB model is

$$\tilde{U}_\theta(\hat{\boldsymbol{\delta}}_0) = \sum_{i=1}^{N} \left\{ (1 - I_{\{y_i=0\}})\hat{\nu}_i + I_{\{y_i=0\}} \binom{n_i}{2} \hat{\pi}_i \hat{\kappa}_i / (1-\hat{\pi}) \right\},$$

where

$$\hat{\nu}_i = \hat{\pi}_i^{-1}\binom{y_i}{2} + (1-\hat{\pi}_i)^{-1}\binom{n_i - y_i}{2} - \binom{n_i}{2}.$$

Again, it is possible to write this test statistic in a more intuitive form as a linear combination of differences between sample moments and the zero frequency and their corresponding theoretical values. Specifically, $\tilde{U}_\theta(\hat{\boldsymbol{\delta}}_0)$ can be written as

$$\frac{1}{2}\sum_{i=1}^{N}\left[A_i\left\{(y_i-\hat{\mu}_i)^2 - \hat{\sigma}_i^2\right\} + B_i(y_i - \hat{\mu}_i) + C_i(I_{\{y_i=0\}} - \hat{\rho}_{0,i})\right], \tag{5}$$

where now

$$\hat{\mu}_i = (1-\hat{p}_i)n_i\hat{\pi}_i, \quad \hat{\sigma}_i^2 = \hat{\mu}_i\{1 - \hat{\pi}_i(1-\hat{p}_i n_i)\}$$

and

$$A_i = \hat{\pi}_i^{-1}(1-\hat{\pi}_i)^{-1}, \quad B_i = \{2\hat{\pi}_i(1-\hat{p}_i n_i) - 1\}/\{\hat{\pi}_i(1-\hat{\pi}_i)\}$$

and $C_i = -n_i(n_i-1)\hat{\pi}_i\hat{p}_i/\{\hat{\rho}_{0,i}(1-\hat{\pi}_i)\}$.

The score test statistic is given by $\tilde{T} = \tilde{U}_\theta(\hat{\boldsymbol{\delta}}_0)\sqrt{\tilde{I}^{\theta\theta}(\hat{\boldsymbol{\delta}}_0)}$, where $\tilde{I}^{\theta\theta}(\hat{\boldsymbol{\delta}}_0)$ is obtained from the information matrix $\tilde{\mathbf{I}}$. This matrix has typical elements

$$\begin{aligned} \tilde{I}_{\theta\theta} &= \sum_{i=1}^{N}\binom{n_i}{2}\left\{(1-p_i) - \binom{n_i}{2}p_i\pi_i^2(1-\pi_i)^{-2}\kappa_i\right\}, \\ \tilde{I}_{\theta\gamma_j} &= \sum_{i=1}^{N}\binom{n_i}{2}\frac{p_i\pi_i\kappa_i G_{ij}}{1-\pi_i}, \\ \tilde{I}_{\theta\beta_j} &= \sum_{i=1}^{N}\binom{n_i}{2}\frac{n_i p_i \pi_i^2\kappa_i B_{ij}}{1-\pi_i}, \end{aligned}$$

$$\tilde{I}_{\gamma_j\gamma_k} = \sum_{i=1}^{N} p_i^2 G_{ij} G_{ik} (\rho_{0,i}^{-1} - 1), \quad (6)$$

$$\tilde{I}_{\gamma_j\beta_k} = -\sum_{i=1}^{N} n_i p_i \pi_i \kappa_i G_{ij} B_{ik}, \quad (7)$$

$$\tilde{I}_{\beta_j\beta_k} = \sum_{i=1}^{N} n_i \pi_i B_{ij} B_{ik} \left\{(1-p_i)(1-\pi_i) - n_i \pi_i p_i \kappa_i\right\}. \quad (8)$$

Under the ZIB-mixed alternative, the score statistic (3) becomes

$$U_\theta(\hat{\boldsymbol{\delta}}_0) = \frac{1}{2} \sum_{i=1}^{N} \left[\left\{(y_i - n_i\hat{\pi}_i)^2 - n_i\hat{\pi}_i(1-\hat{\pi}_i)\right\} - I_{\{y_i=0\}} n_i\hat{\pi}_i(n_i\hat{\pi}_i + \hat{\pi}_i - 1)\hat{p}_i/\hat{\rho}_{0,i}\right].$$

An alternative representation is available that has the form given in (5), but with

$$A_i = 1, \quad B_i = -2\hat{p}_i n_i \hat{\pi}_i \quad \text{and} \quad C_i = n_i\hat{\pi}_i\hat{p}_i(1 - \hat{\pi}_i - n_i\hat{\pi}_i)/\hat{\rho}_{0,i}.$$

The score test statistic is $T = U_\theta(\hat{\boldsymbol{\delta}}_0)\sqrt{I^{\theta\theta}(\hat{\boldsymbol{\delta}}_0)}$, where $I^{\theta\theta}(\hat{\boldsymbol{\delta}}_0)$ is obtained from the information matrix **I** given in general form in the Appendix. In this non-clustered special case, formulas for the elements of **I** reduce to

$$I_{\theta\theta} = \frac{1}{4} \sum_{i=1}^{N} \left[(1-p_i)n_i\pi_i(1-\pi_i)\{1 - 2\pi_i(1-\pi_i)(3-n_i)\} - p_i\kappa_i\{n_i\pi_i(n_i\pi_i + \pi_i - 1)\}^2\right],$$

$$I_{\theta\gamma_j} = \frac{1}{2} \sum_{i=1}^{N} p_i n_i \pi_i (n_i\pi_i + \pi_i - 1)\kappa_i G_{ij},$$

$$I_{\theta\beta_j} = \frac{1}{2} \sum_{i=1}^{N} n_i\pi_i B_{ij} \left\{p_i n_i \pi_i \kappa_i (n_i\pi_i + \pi_i - 1) - (1-p_i)(2\pi_i - 1)(1-\pi_i)\right\},$$

and $I_{\gamma_j\gamma_k}$, $I_{\gamma_j\beta_k}$ and $I_{\beta_j\beta_k}$ are as given in (6), (7), and (8) for the ZIBB alternative.

Note that our proposed score test statistics T and $\tilde{T}$ for overdispersion relative to a ZIB model generalize the tests of Dean (1992) for extra-binomial variation. That is, for no zero-inflation (i.e., $p_{ij} = 0$ for all i, j), T and $\tilde{T}$ reduce to the test statistics N_A and N_B, respectively, in the notation of Dean's paper.

5. EXAMPLES

5.1. Testing for overdispersion in a ZIB model.

To illustrate our ZIB overdispersion tests, we consider a data set concerning alligator egg hatch rates. The data were collected by researchers from the Florida Fish and Wildlife Conservation Commission and the Florida Cooperative Fish and Wildlife Research Unit, U. S. Geological Survey. The data, summarized in Table 1, consist of hatch counts and clutch sizes for 109 clutches of alligator eggs laid during the 1988 through 1991 breeding seasons at Lake George, Florida.

Let y_{ij} be the number of hatched eggs out of n_{ij}, the clutch size, for the jth clutch during the ith year. Based on a simple binomial logistic regression model with mean specification $\text{logit}(\pi_{ij}) = \beta_i$, the score test of Deng & Paul (2000) for zero-inflation indicates overwhelming evidence of excess zeros relative to a binomial model (test statistic $= 624, 330.6$). Therefore, we considered a ZIB model for these data with the same linear predictor for π_{ij} and $\text{logit}(p_{ij}) = \gamma_i$, for the mixing probability. Although it is clearly necessary to account for zero-inflation for these data, the question remains whether there is overdispersion relative to a ZIB model. That is, would either a ZIB-mixed or ZIBB model be more appropriate? Based on the ZIB null model, the tests of Section 4.2 all give very similar and highly significant results. For the ZIBB alternative, we obtain $\tilde{T} = 70.9$ and for the ZIB-mixed aternative, we get $T = 70.6$ with a small-sample corrected value of $T^c = 70.9$.

TABLE 1: Summary of alligator egg viability data.

Year	Number of clutches	Average clutch size	Percent hatched	Number of zeros
1988	22	43.5	43.2	3
1989	23	46.2	27.2	4
1990	31	43.4	65.1	0
1991	33	45.9	56.5	3

These results strongly indicate overdispersion relative to a ZIB model for these data. To further investigate the suitability of a ZIB-mixed model versus a ZIB model in this example, we follow the approach of Vieira, Hinde & Demétrio (2000), who suggested the use of half-normal plots as goodness-of-fit tools for models for zero-inflated proportion data. Half-normal plots for the ZIB model described above and for a similar ZIB-mixed model with

$$\text{logit}(\pi_{ij}) = \beta_i + b_{ij}, \quad \text{logit}(p_{ij}) = \gamma_i \tag{9}$$

appear in Figure 1. The plots display the absolute values of the Pearson residuals versus half-normal scores, with simulated envelopes based on the assumed model evaluated at the estimated parameter values. A suitable model is indicated by the observed values falling within the simulated envelope. The Pearson residuals in this context are defined as

$$\left\{y_{ij} - \widehat{\text{E}(y_{ij})}\right\} / \widehat{\text{var}^{1/2}(y_{ij})},$$

where

$$\begin{aligned}
\text{E}(y_{ij}) &= \text{E}\{\text{E}(y_{ij}|b_i)\}, \\
\text{var}(y_{ij}) &= \text{E}\{\text{var}(y_{ij}|b_i)\} + \text{E}[\{\text{E}(y_{ij}|b_i)\}^2] - \{\text{E}(y_{ij})\}^2, \\
\text{E}(y_{ij}|b_i) &= (1-p_{ij})n_{ij}\pi_{ij}, \\
\text{var}(y_{ij}|b_i) &= (1-p_{ij})n_{ij}\pi_{ij}\{1-\pi_{ij}(1-p_{ij}n_{ij})\}.
\end{aligned}$$

The marginal expectations here (necessary only for the ZIB-mixed model) were evaluated using 9-point Gaussian quadrature and the hats indicate evaluation at the final parameter estimates. Figure 1(a) confirms the results of the overdispersion tests indicating that the ZIB model is inadequate for these data. In addition, Figure 1(b) suggests that a ZIB-mixed model fits reasonably well, and provides an appropriate alternative hypothesis for testing overdispersion relative to the ZIB null model.

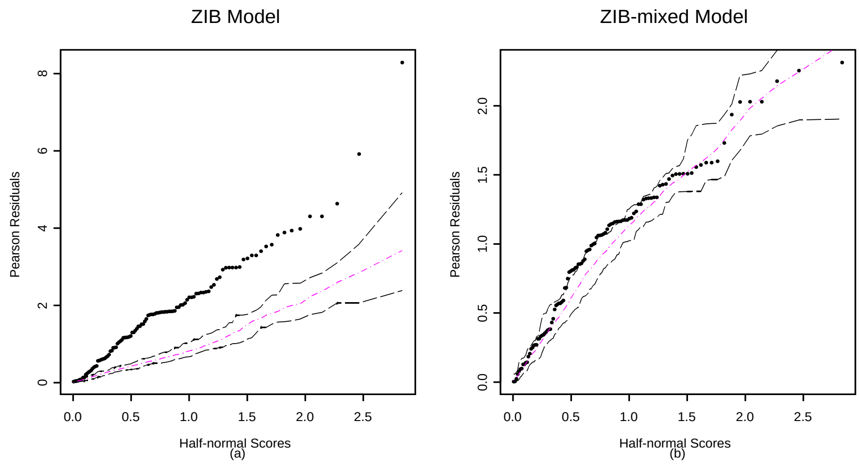

FIGURE 1: Half-normal plots for ZIB and ZIB-mixed models for alligator egg data.

5.2. Testing vector-valued variance components in a ZIB model.

To illustrate our general test introduced in Section 3, we consider a horticultural data set described and analyzed in van Iersel, Oetting, Hall & Kang (2001). The data come from an experimental design that can be described as a randomized complete block design with pseudo-replication (sub-sampling) and repeated measures. Fifteen units of three plants each were randomized into five complete blocks and three insecticide treatments were applied. These 15 units were then followed over time with measurements (pseudo-replicates) taken in two locations on two plants per unit at seven occasions in time. Note that at each measurement occasion, only two of the three plants in each unit were randomly selected for measurement, with each selected plant measured in two locations (low and high) on the plant. A similar data set from a slightly less complex design was used to motivate and describe ZIP-mixed and ZIB-mixed models in Hall (2000). In that case, one observation per plant was obtained at each of several measurement occasions, yielding a repeated measures design. For the data we consider here, the design is a doubly-repeated measures design in the sense that the response was measured twice (at two locations) for each plant at each of several measurement occasions through time.

One of the response variables measured in the study was a bounded count, the number of surviving insects from a known number (usually about 10) of insects placed in a specific location on the plant three days prior to measurement. Because this response variable exhibits many more zeros than would be expected according to a binomial distribution, van Iersel, Oetting, Hall & Kang (2001) analyzed the data using ZIB-mixed models. The doubly-repeated nature of the design suggests that one might account for the within-cluster correlation that is to be expected among multiple measurements on each plant by including plant specific random effects in a ZIB model. In particular, van Iersel, Oetting, Hall & Kang (2001) considered ZIB-mixed models of the form described in Section 2.2, with each plant taken to represent a cluster on which repeated measures were taken, and where bivariate cluster specific random effects were included in the model.

Let y_{ij}/n_{ij} represent the proportion of surviving insects for the jth measurement taken on the ith plant, where $i = 1, \ldots, K$ ($K = 30$) and $j = 1, \ldots, t_i$. Here t_i, the number of responses available on the ith plant, would have been (# measurement locations) $\times$ (# measurement occasions) $= (2)(7) = 14$ if each plant had been measured at all measurement occasions. However, because of the random sampling of two of three plants per unit at each measurement occasion, t_i varied from plant to plant. The ZIB-mixed models considered for these data by van Iersel, Oetting, Hall & Kang (2001) assumed that $\boldsymbol{\pi}_i$, the vector of survival probabilities for the t_i responses on the ith plant, is related to covariates and random effects via

$$\text{logit}(\boldsymbol{\pi}_i) = \mathbf{B}_i\boldsymbol{\beta} + b_{1i} + b_{2i}\mathbf{loc}_i, \quad i = 1, \ldots, K$$

where $\mathbf{loc}_i$ is a $t_i \times 1$ vector of indicators for whether or not the (i, j)th response was taken at location 1. Here we have allowed for plant-to-plant heterogeneity through b_{1i} and heterogeneity in the location effect from plant to plant through b_{2i}. Models of this form were fitted under the assumption that $\mathbf{b}_i = (b_{1i}, b_{2i})'$, $i = 1, \ldots, K$, were independent, bivariate normal random vectors with mean zero and nondiagonal variance-covariance matrix,

$$\mathbf{D} = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_3 \end{pmatrix}. \tag{10}$$

Here we consider testing $\mathcal{H}_0 : \boldsymbol{\theta} = \mathbf{0}$ using the score test of Section 3.

For illustration, we performed this test using a somewhat simpler model than that used as the basis of final analysis in van Iersel, Oetting, Hall & Kang (2001). Specifically, we assume linear predictors $\mathbf{B}_i\boldsymbol{\beta}$ and $\mathbf{G}_i\boldsymbol{\gamma}$ of the same form, each including main effects for blocks, treatments, locations on the plants, and measurement occasions. The resulting score test statistic for H_0 is $T^2 = 13.46$, which yields an approximate $p-$value of 0.0037 based on an asymptotic $\chi^2(3)$ distribution under the null hypothesis. The small-sample adjusted test statistic is $(T^c)^2 = 16.14$ with asymptotic p-value $p = 0.0011$. The maximized log-likelihoods for this model with and without the presence of the random effects are -643.19 and -653.26, respectively, so a likelihood ratio test statistic for $\mathcal{H}_0$ has value $2(-643.19 + 653.26) = 20.13$. It is well known, however, that the chi-square limiting distribution for the likelihood ratio test no longer applies in this context in which the null hypothesis places the parameter on the boundary of its parameter space. Instead, the likelihood ratio statistic is asymptotically distributed as a complex mixture of chi-squares from which a $p-$value is not easily obtained.

For illustration, we also consider testing homogeneity in the ZIB-mixed model with the same linear predictors $\mathbf{B}_i\boldsymbol{\beta}$ and $\mathbf{G}_i\boldsymbol{\gamma}$ as before, but now with a single random effect at the plant level. That is, we consider testing $\mathcal{H}_0 : \theta = 0$ in the ZIB-mixed model with $\text{logit}(\boldsymbol{\pi}_i) = \mathbf{B}_i\boldsymbol{\beta} + b_i$, and

$$\text{logit}(\mathbf{p}_i) = \mathbf{G}_i\boldsymbol{\gamma}, \quad i = 1, \ldots, K, \quad \text{with } b_1, \ldots, b_K \overset{\text{iid}}{\sim} N(0, \theta).$$

The score test statistic for $\mathcal{H}_0$ is $T = 1.91$, which yields a one-sided p-value of $p = .0284$. The small-sample corrected version of this test statistic is $T^c = 2.20$ ($p = .0139$). Note that the alternative hypothesis model here is similar to, but distinct from the alternative hypothesis model for the overdispersion test T of Section 4.2. For the overdispersion test, the model under the alternative hypothesis has $\text{logit}(\pi_{ij}) = \mathbf{B}_{ij}\boldsymbol{\beta} + b_{ij}$, where $\mathbf{B}_{ij}$ is the jth row of $\mathbf{B}_i$ and

$$b_{11}, \ldots, b_{K,n_K} \overset{\text{iid}}{\sim} N(0, \theta).$$

To distinguish between the univariate and bivariate random effects model considered above, it would be useful to have a test for the null hypothesis $\mathcal{H}_0 : \theta_1 \in [0, \infty), \theta_2 = \theta_3 = 0$ in (10). Score tests for such composite hypotheses on $\boldsymbol{\theta}$ have not yet been studied. Lin (1997) has considered score tests for the problem of testing individual variance components equal to zero in generalized linear mixed effects models with independent normal random effects, and her approach could be adapted to zero-inflated regression models. More directly applicable is the work of Stram & Lee (1994), who consider likelihood ratio tests for variance components in linear mixed effects models. Their results can be expected to apply in our context and lead to the conclusion that the likelihood ratio test statistic for $\mathcal{H}_0$ has an asymptotic distribution that is a 50:50 mixture of χ^2_1 and χ^2_2. Therefore, comparison to the naive reference distribution of χ^2_2 leads to a conservative test. In our example, the maximized log-likelihood for the univariate random effects ZIB-mixed model is -647.66, so that the likelihood ratio test statistic is $2(-643.19 + 647.66) = 8.94$. Comparison against the χ^2_2 distribution yields a conservative $p-$value of $p = .0114$, so we reject $\mathcal{H}_0$ and conclude that bivariate random effects are necessary here. Less formally, we reach the same conclusion by noting that the score test of homogeneity was more significant for the bivariate random effects alternative model ($p = .0011$ for the small-sample adjusted test T^c) than for the univariate random effects alternative ($p = .0139$ for T^c).

6. A SIMULATION STUDY OF POWER AND SIZE

We now present the results of a small simulation study examining the empirical size and power of the test statistics discussed in this paper. Here we limit our attention to the overdispersion tests of Section 4. Two sets of simulations were run: one based on the alligator egg viability data set discussed in Section 5.1 to investigate the properties of the ZIB overdispersion tests and a second set to examine the ZIP overdispersion tests. The latter set of simulations is based on data that were presented by Ridout, Demétrio & Hinde (2001) to illustrate their score test for overdispersion in a ZIP model.

The first set of simulations was based on model (9). One thousand sets of data were generated from this model with $\boldsymbol{\beta} = (0.56, 0.00, -0.72, 0.66)'$ and $\boldsymbol{\gamma} = (-2.30, -1.85, -1.56, -11.34)'$. These values correspond to the parameter estimates when the model was fit to the actual data. These parameter values imply that the vector of hatching probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)'$ and the vector of zero-inflation probabilities $\mathbf{p} = (p_1, p_2, p_3, p_4)'$ for the four years of the study were $\boldsymbol{\pi} = (0.64, 0.50, 0.33, 0.66)'$ and $\mathbf{p} = (0.091, 0.14, 0.17, 0.000012)'$, respectively. We considered three values of n, the number of clutches: $n = 20$, $n = 60$, and $n = 100$; and we examined five values of θ, the variance component associated with the random effect b_i in model (9): $\theta = 0$, $\theta = (0.10)^2$, $\theta = (0.25)^2$, $\theta = (0.50)^2$, and $\theta = (0.75)^2$. The binomial denominators, n_i, $i = 1, \ldots, n$ (the clutch sizes), were generated from a $N(45, 81)$ distribution, which corresponds to the sample mean and variance of the clutch sizes in the original data set. Results are displayed in Table 2. It is clear from these results that the size ($\theta = 0$) and power ($\theta > 0$) of the unadjusted tests T and $\tilde{T}$ are very similar and substantially deflated in comparison to the small-sample adjusted test T^c. The size of T^c is reasonably close to nominal (5%) even in the very small sample case $n = 20$. We conclude that the chi-square approximation to the distribution of the unadjusted tests is quite poor in small sample sizes and that the adjusted test should be used instead.

The second set of simulations is based on data from Ridout, Demétrio & Hinde (2001). These authors illustrate their score test for testing a ZIP model versus a ZINB alternative using data on the number of roots produced by 270 shoots of a certain apple cultivar. The shoots were produced under a 2×4 factorial treatment structure, where the treatment factors were photoperiod (8 or 16 hours) and cytokin concentration (2.2, 4.4, 8.8, or 17.6 μM). In the original data set, either 30 or 40 shoots were present in each treatment combination.

Let y_{ijk} be the number of roots on the kth shoot at the ith level of photoperiod and jth level of cytokin concentration, and let $\lambda_{ijk} = \mathrm{E}(y_{ijk})$. Then the model on which we based our simulations is the same as that considered by Ridout, Demétrio & Hinde (2001), namely:

$$\log(\lambda_{ijk}) = \beta_{ij}, \quad \mathrm{logit}(p_{ij}) = \gamma_i.$$

We generated 1000 data sets from this model with n_{ij}, the number of shoots in the (i, j)th treatment equal to 4, 6, and 12, for total sample sizes of $n = 32$, $n = 48$ and $n = 96$, respectively. Again, the parameter values under which the data were generated were set equal to the fitted values from the original data set. These values were $\boldsymbol{\beta} = (1.76, 2.05, 2.01, 2.02, 1.88, 1.76, 1.65, 1.53)'$ and $\boldsymbol{\gamma} = (-4.27, -0.10)'$. In Table 3, we present results for our test T and the corresponding small-sample adjusted version T^c. In addition, we present results for the test of Ridout, Demétrio & Hinde (2001) corresponding to the $c = 2$ parameterization of the negative binomial distribution. This test is labelled R_2 in Table 3. As mentioned in Section 4.1, the test of Ridout, Demétrio & Hinde under the $c = 1$ parameterization, denoted R_1 in Table 3, coincides with our test T. The same general pattern of results can be seen in Tables 2 and 3. The small-sample correction is clearly necessary for the chi-square approximation to be appropriate for sample sizes in the range considered in these simulations.

TABLE 2: Percentage of rejections of $\mathcal{H}_0$: ZIB model holds, for data generated from a ZIB-mixed model with variance component θ. Results based on 1000 replications.

	Test			θ		
n	statistic	0^2	$(.10)^2$	$(.25)^2$	$(.50)^2$	$(.75)^2$
	$\tilde{T}$	1.40	3.30	33.40	88.90	98.90
20	T	1.40	3.10	33.30	88.90	98.90
	T^c	4.40	10.30	47.60	93.20	99.40
	$\tilde{T}$	2.80	5.90	60.10	100.00	100.00
60	T	2.80	6.10	59.70	100.00	100.00
	T^c	5.70	12.50	70.20	100.00	100.00
	$\tilde{T}$	3.80	14.30	93.20	100.00	100.00
100	T	3.80	14.30	93.20	100.00	100.00
	T^c	6.40	20.10	95.10	100.00	100.00

TABLE 3: Percentage of rejections of $\mathcal{H}_0$: ZIP model holds, for data generated from a ZIP-mixed model with variance component θ. Results based on 1000 replications.

	Test			θ		
n	statistic	0^2	$(.10)^2$	$(.25)^2$	$(.50)^2$	$(.75)^2$
	R_2	0.20	0.20	6.89	67.66	95.21
32	$R_1 = T$	0.40	0.80	11.08	72.55	96.01
	T^c	3.01	6.29	27.64	87.23	98.40
	R_2	1.00	1.80	15.90	91.30	99.50
48	$R_1 = T$	1.40	2.30	18.80	92.10	99.60
	T^c	5.90	7.30	38.20	97.20	99.70
	R_2	0.50	3.60	46.60	100.00	100.00
96	$R_1 = T$	0.60	4.80	48.40	100.00	100.00
	T^c	5.00	11.10	68.00	100.00	100.00

7. DISCUSSION

In this paper, we have proposed a score test for testing that variance and covariance components are zero in a ZIP-mixed or ZIB-mixed regression model. A score test for this problem is particularly convenient for two reasons: first, because it only necessitates fitting the model under the null hypothesis, which is substantially easier than fitting the mixed-effects alternative model; and second, because unlike the Wald and likelihood ratio tests, the score test retains its asymptotic chi-square distribution in this nonstandard testing problem in which the null hypothesis places the parameter on the boundary of its parameter space.

In the overdispersion problem of Section 4 in which θ is a scalar, the score test is easily made into a one-sided test by rejecting at significance level α for T exceeding the upper αth quantile of the standard normal distribution. A one-sided alternative $H_A : \theta > 0$ is clearly appropriate in this situation in which θ is a variance component. In the case in which $\dim(\theta) > 1$, forming a one-sided test is not quite as simple, although it is equally appropriate. The appropriate alternative in the nonscalar case is that $\boldsymbol{\theta}$ be constrained so that $\mathbf{D}(\boldsymbol{\theta})$ is positive definite. Hall & Præstgaard (2001) have shown how the score test may be altered for this alternative hypothesis by first projecting the score vector onto the tangent cone of the model before forming the test statistic. The resulting test is more powerful in the same way that the one-sided test is more powerful than the two-sided test when $\boldsymbol{\theta}$ is known to be non-negative. Unfortunately, the projected score test statistic no longer follows a simple chi-square limiting distribution. Instead, its asymptotic distribution is a mixture of chi-squares. For simplicity, we have chosen to present the nonprojected test in this paper. However, the modification of Hall & Præstgaard (2001) is available in this context and does offer greater power.

APPENDIX

Here we present formulas for the information matrix in the general case. Let $a^j_{i,\ell\ell'}$ denote the (ℓ,ℓ')th element of $\mathbf{W}_i\dot{\mathbf{D}}_j\mathbf{W}_i$. Then after some calculations similar to those described in Lin (1997, Appendix 1), we obtain the (j,k)th component of $\mathbf{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ as

$$I_{\theta_j\theta_k} = \frac{1}{4}\sum_{i=1}^{K}\left(\sum_{\ell=1}^{t_i} a^j_{i,\ell\ell}a^k_{i,\ell\ell}r_{i,\ell\ell} + 4\sum_{\ell<\ell'} a^j_{i,\ell\ell'}a^k_{i,\ell\ell'}r_{i,\ell\ell'}\right),$$

where $r_{i,\ell\ell} = -p_{i\ell}(1-p_{i\ell}/\rho_{0,i\ell})\left[\{b'(\eta_{i\ell})\}^2 - b''(\eta_{i\ell})\right]^2 + (1-p_{i\ell})[b''''(\eta_{i\ell}) + 2\{b''(\eta_{i\ell}\}^2]$ and

$$\begin{aligned} r_{i,\ell\ell'} &= \left[(1-p_{i\ell})b''(\eta_{i\ell}) - \{b'(\eta_{i\ell})\}^2 p_{i\ell}(1-p_{i\ell}/\rho_{0,i\ell})\right] \\ &\quad \times \left[(1-p_{i\ell'})b''(\eta_{i\ell'}) - \{b'(\eta_{i\ell'})\}^2 p_{i\ell'}(1-p_{i\ell'}/\rho_{0,i\ell'})\right]. \end{aligned}$$

In the general case, the components of the score vector at $\mathcal{H}_0$ corresponding to $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are given by

$$U_{\gamma_k} = \sum_{i=1}^{K}\sum_{j=1}^{t_i} p_{ij}G_{ij,k}\left(I_{\{y_{ij}=0\}}/\rho_{0,ij} - 1\right), \quad k = 1,\ldots,r,$$

and

$$U_{\beta_k} = \sum_{i=1}^{K}\sum_{j=1}^{t_i} B_{ij,k}\left\{y_{ij} - b'(\eta_{ij}) + I_{\{y_{ij}=0\}}b'(\eta_{ij})p_{ij}/\rho_{0,ij}\right\}, \quad k = 1,\ldots,p.$$

Using these expressions, we can show that

$$I_{\theta_j\beta_k} = \frac{1}{2}\sum_{i=1}^{K}\sum_{\ell=1}^{t_i} a^j_{i,\ell\ell}B_{i\ell,k}\left((1-p_{i\ell})b'''(\eta_{i\ell}) + p_{i\ell}b'(\eta_{i\ell})\kappa_{i\ell}\left[\{b'(\eta_{i\ell})\}^2 - b''(\eta_{i\ell})\right]\right),$$

$$
\begin{aligned}
I_{\beta_j\beta_k} &= \sum_{i=1}^{K}\sum_{\ell=1}^{t_i} B_{i\ell,j}B_{i\ell,k}\left[(1-p_{i\ell})b''(\eta_{i\ell}) - p_{i\ell}\left\{b'(\eta_{i\ell})\right\}^2\kappa_{i\ell}\right],\\
I_{\theta_j\gamma_k} &= \frac{1}{2}\sum_{i=1}^{K}\sum_{\ell=1}^{t_i} a^j_{i,\ell\ell}p_{i\ell}G_{i\ell,k}\kappa_{i\ell}\left[\left\{b'(\eta_{i\ell})\right\}^2 - b''(\eta_{i\ell})\right],\\
I_{\gamma_j\gamma_k} &= \sum_{i=1}^{K}\sum_{\ell=1}^{t_i} p^2_{i\ell}G_{i\ell,j}G_{i\ell,k}(\rho^{-1}_{0,i\ell}-1), \quad \text{and}\\
I_{\gamma_j\beta_k} &= -\sum_{i=1}^{K}\sum_{\ell=1}^{t_i} p_{i\ell}G_{i\ell,j}B_{i\ell,k}b'(\eta_{i\ell})\kappa_{i\ell},
\end{aligned}
$$

where $\kappa_{i\ell} = 1 - p_{i\ell}/\rho_{0,i\ell}$.

ACKNOWLEDGEMENTS

The authors wish to extend their sincere thanks to the Florida Fish and Wildlife Conservation Commission and the Florida Cooperative Fish and Wildlife Research Unit, U. S. Geological Survey, for kindly providing the alligator egg viability data of Section 5.1.

REFERENCES

D. R. Cox (1983). Some remarks on overdispersion. *Biometrika*, 70, 269–274.

C. B. Dean (1992). Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87, 451–457.

C. Dean & J. F. Lawless (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, 84, 467–472.

D. Deng & S. R. Paul (2000). Score tests for zero-inflation in generalized linear models. *The Canadian Journal of Statistics*, 28, 563–570.

D. B. Hall (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56, 1030–1039.

D. B. Hall & J. T. Præstgaard (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, 88, 739–751.

X. Lin (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84, 309–326.

J. Mullahy (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.

J. C. Pinheiro & D. M Bates (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, 4, 12–35.

J. C. Pinheiro & D. M. Bates (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, 6, 289–296.

R. L. Prentice (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81, 321–327.

M. Ridout, C. G. B. Demétrio & J. Hinde (1998). Models for count data with many zeros. Invited Paper, The XIXth International Biometric Conference, Cape Town, South Africa, pp. 179–192.

M. Ridout, C. G. B. Demétrio & J. Hinde (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57, 219–223.

D. O. Stram & J. W. Lee (1994). Variance component testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171–1177.

J. van den Broek (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, 51, 738–743.

M. van Iersel, R. Oetting, D. B. Hall & J. G. Kang (2001). Application technique and watering method affect imidacloprid control of whiteflies (*Homoptera: Aleyrodidæ*) on poinsettias. *Journal of Economic Entomology*, 94, 666–672.

A. M. C. Vieira, J. P. Hinde & C. G. B. Demétrio (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, 27, 373–389.

R. D. Wolfinger (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. SAS User's Group International Paper #287.

Received 9 July 2001
Accepted 12 February 2002

Daniel B. HALL: dhall@stat.uga.edu
Department of Statistics, University of Georgia
Athens, GA 30602-1952, USA

Kenneth S. BERENHAUT: kberenht@mthcsc.wfu.edu
Department of Mathematics, Wake Forest University
P. O. Box 7388, Winston-Salem, NC 27109, USA