Journal of Applied Statistics

Multivariate models for correlated count data

Mariana Rodrigues-Motta, Hildete P. Pinheiro, Eduardo G. Martins, Márcio S. Araújo & Sérgio F. dos Reis

Department of Statistics, University of Campinas, Campinas, 13083-859, Brazil
Department of Forest Sciences, Centre for Applied Conservation Research, University of British Columbia, Vancouver, Canada, V6T1Z4
Department of Biology, Institute of Environmental Science, Carleton University, Ottawa, Canada, K1S5B6
Departamento de Ecologia, Universidade Estadual Paulista, Rio Claro, 13506-900, Brazil
Departamento de Biologia Animal, Universidade Estadual de Campinas, Campinas, 13083-862, Brazil

Published online: 14 Apr 2013.


To link to this article: http://dx.doi.org/10.1080/02664763.2013.789098

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the “Content”) contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.
Multivariate models for correlated count data

Mariana Rodrigues-Motta*, Hildete P. Pinheiroa, Eduardo G. Martinsb,c, Márcio S. Araújod and Sérgio F. dos Reise

aDepartment of Statistics, University of Campinas, Campinas 13083-859, Brazil; bDepartment of Forest Sciences, Centre for Applied Conservation Research, University of British Columbia, Vancouver, Canada V6T1Z4; cDepartment of Biology, Institute of Environmental Science, Carleton University, Ottawa, Canada K1S5B6; dDepartamento de Ecologia, Universidade Estadual Paulista, Rio Claro 13506-900, Brazil; eDepartamento de Biologia Animal, Universidade Estadual de Campinas, Campinas 13083-862, Brazil

(Received 18 January 2012; accepted 20 March 2013)

In this study, we deal with the problem of overdispersion beyond extra zeros for a collection of counts that can be correlated. Poisson, negative binomial, zero-inflated Poisson and zero-inflated negative binomial distributions have been considered. First, we propose a multivariate count model in which all counts follow the same distribution and are correlated. Then we extend this model in a sense that correlated counts may follow different distributions. To accommodate correlation among counts, we have considered correlated random effects for each individual in the mean structure, thus inducing dependency among common observations to an individual. The method is applied to real data to investigate variation in food resources use in a species of marsupial in a locality of the Brazilian Cerrado biome.

Keywords: maximum likelihood; mixed model; mixture distribution; multivariate count data; overdispersion; Poisson distribution; negative binomial distribution; zero-inflated data

1. Introduction

The dynamics of food resources use in natural populations is determined by complex interactions between endogenous (e.g. sex, age and morphology) and environmental variables which determine fluctuations in resource availability [15,29–31,34,35]. As remarked by [3], intrapopulation variation in food resource use is best evaluated within the framework of statistical linear models, and counts of food items consumed by individuals in a population provide the necessary raw data to model variation in resource use. For counts of food items, generalized linear models are the appropriate statistical formalism with the standard Poisson regression model as a natural choice [1,2,6,8,12,22,23]. Nevertheless, diet data are inherently spatial and temporal and, therefore, heterogeneous and very frequently violate the Poisson model hallmark assumption of equidispersion,
i.e. mean = variance [23]. The violation of this assumption, often caused by overdispersion in the data, occurs when the observed sample variance is larger than that given by substituting the sample mean into the theoretical mean–variance relationship [40]. This overdispersion in turn may stem from either large counts or an excess of zeros in the data [19,27].

A well-established implication of overdispersed data sets is that statistical inference can be biased [23]. As a consequence, there has been recently a strong interest in the literature in the application of alternative models of increasing complexity that can handle the common properties of ecological data, which often violate the equidispersion assumption of the standard Poisson regression model [18,27,36]. These efforts have been aimed primarily at the estimation of parameters and associated measures of uncertainty from ecological data describing commonness and rarity of species, patch occupancy and metapopulation dynamics, recruitment, and mating success [9,10,17,33,39,44]. Nevertheless, the extent to which diet data resulting from complex interactions between individuals and their environment deviate from the equidispersion expectation of the Poisson model apparently has not yet been examined.

In this article, we propose a multivariate model for correlated count data with the presence of overdispersion. Poisson, negative binomial (NB), zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) distributions have been considered in this study [1,4,38,41,43]. The Poisson model is proposed as the primary choice while the NB model is an option when overdispersion is present. To handle extra zeros the ZIP and the ZINB models are considered; the ZIP and ZINB models incorporate more zeros then allowed by the Poisson and NB, respectively. To account for correlation among counts, a collection of random effects is used, thus inducing a marginal dependency among observations common to an individual. An extension of this multivariate model is considered by allowing each count to follow one of the four proposed models, handling the presence of overdispersion and/or the presence of extra zeros differently among the correlated counts. For all models, we employ a generalized linear model framework [23], and covariates affecting the mean are allowed to differ across counts. Conditioned on the random effects, counts are assumed to be independent with a conditional mean function that depends on the random effects and a set of covariates. Model fitting is checked by using simulated envelopes comparing standard Pearson residuals with half-normal scores. Finally, in Section 3 the methodology is applied to investigate intrapopulation variation in food resources use in the didelphid marsupial *Gra
cilinanus microtarsus* [4,19–21] in a locality of the Brazilian Cerrado biome, a hotspot for biodiversity [32].

2. Statistical methodology

2.1 Multivariate models for count data

Let \( \mathbf{y}_i = (y_{i1}, \ldots, y_{id})^T \) denote the collection of \( J \) counts on subject \( i \), where \( i = 1, \ldots, n \). Let all data be represented by \( \mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T) \). Let \( \mathbf{u} = (u_{i1}, \ldots, u_{id})^T \) denote a set of \( J \) subject and outcome-specific random effects, and suppose that, conditioned on \( \mathbf{u}_i \), the counts \( y_{ij} \) are independent. To model correlation among counts, let \( \mathbf{u}_i \sim N_J(\mathbf{0}, \mathbf{D}) \), where \( \mathbf{D} \) is an unrestricted covariance matrix [7]. Following [28] we assume that, conditioned on a random effect \( \mathbf{u}_i \), the conditional expectation is modeled in a logarithm scale as

\[
\log(E(y_{ij} | \mathbf{u}_i)) = \log(\theta_{ij}) = \mathbf{x}_{ij}^T \bm{\beta}_j + u_{ij},
\]

where \( \mathbf{x}_{ij}^T \) is a row vector of covariates not necessarily the same for all \( j \) and \( \bm{\beta}_j \in \mathbb{R}^d_j \) is the vector of parameters.
The natural choice for fitting a count data type is the Poisson model, but in order to explore the possibility of extra-Poisson variation, we use the NB regression. The NB distribution model accounts for overdispersion in the dependent variable by defining the variance as a function of the mean [23]. Neither the Poisson nor the NB regression models account for the presence of an excess of zeros [18, 27]. The models ZIP and ZINB are used to handle extra zeros. In zero-inflated count models, zeros may come from the ‘perfect state’ with probability \( p \) and from the count component with probability \( 1 - p \), thus the probability of observing a zero is the sum of the probability of those events. Following [42], if \( y_{ij} | u_i \sim \text{ZIP}(p, \theta_j) \) then

\[
y_{ij} | u_i \sim \begin{cases} 
p + (1 - p) e^{-\theta_j}, & y_{ij} = 0, \\
(1 - p) e^{-\theta_j} \frac{\theta_j^{y_{ij}}}{y_{ij}!}, & y_{ij} > 0,
\end{cases}
\]

where \( \theta_j \) has been modeled as Equation (1) with a log-linear regression model. In the ZIP model, \( E(\psi | u_i) = \theta_j(1 - p) \) and \( \text{Var}(\psi | u_i) = E(\psi | u_i)(1 + p\theta_j) \). Now, if \( y_{ij} | u_i \sim \text{ZINB}(p, \theta_j, \kappa) \), its probability distribution function is

\[
y_{ij} | u_i \sim \begin{cases} 
p + (1 - p)(1 + \kappa\theta_j)^{-1/\kappa}, & y_{ij} = 0 \\
(1 - p) (1 + \kappa\theta_j)^{-1} \left( 1 + \frac{1}{\theta_j\kappa} \right)^{-y_{ij}}, & y_{ij} > 0,
\end{cases}
\]

where \( \kappa > 0 \) is the dispersion parameter and an estimate of \( \kappa \) larger than 0 indicates that there exists overdispersion in the data set. Under this model, \( E(\psi | u_i) = \theta_j(1 - p) \) and \( \text{Var}(\psi | u_i) = E(\psi | u_i)(A + p\theta_j) \), with \( A = (1 + \kappa E(\psi | u_i)) \).

Note that the ZINB distribution in Equation (3) reduces to the ZIP distribution in Equation (2), as well as the NB distribution reduces to the Poisson distribution when \( \kappa \to 0 \). In the zero-inflated distributions (2) and (3), \( p \) is the probability of extra zeros and represents the probability of being in the ‘perfect state’. The ZIP and ZINB models reduce to the Poisson and NB, respectively, if \( p = 0 \). Either \( \kappa \) or \( p \) in Equations (2) and (3) is assumed homogeneous among individuals [28, 42].

Let \( \psi = (\beta_0^T, \sigma^2, p, \kappa)^T \) be the combined parameter vector according to the model being estimated. The likelihood function is the product of the contributions \( p(y_{ij} ; \psi) \), where \( p(y_{ij} ; \psi) \) is the joint probability distribution of the \( J \) counts. Therefore, the log-likelihood for the appropriate parameter vector \( \psi \) based on all data is given by

\[
l(\psi ; y, u) = \sum_{i=1}^{n} \log[p(y_{ij} ; \psi)] = \sum_{i=1}^{n} \log \left[ \int \prod_{j=1}^{J} p(y_{ij} | u_{ij}; \psi) \varphi_j(u_i; D) \, du_i \right] \]

where \( p(y_{ij} | u_{ij}; \psi) \) is a Poisson, NB, ZIP or ZINB distribution and \( \varphi_j \) denotes the \( J \)-variate normal density function with unrestricted covariance matrix \( D \) [7]. Maximization of Equation (4) with respect to \( \psi \) is complicated by the integration with respect to \( u_i \). Using \( m \)-point Gaussian quadrature to approximate these integrals [25], we have

\[
l(\psi ; y, u) \approx \sum_{i=1}^{n} \log \left\{ \sum_{l_1=1}^{m} \cdots \sum_{l_J=1}^{m} [p(y_{ij} | u_{i1}; \psi) \cdots p(y_{ij} | u_{iJ}; \psi)] w_{l_1} \cdots w_{l_J} \right\}.
\]
where $u_{ij}$ are quadrature points and $w_i$ the associated weights. The dual quasi-Newton optimization algorithm [25] is used to carry out the maximization of Equation (5) and random effects $u_{ij}$ are predicted using empirical Bayes estimates [25], denoted here by $\hat{u}_{ij}$.

### 2.2 Multivariate extension of the likelihood

The model presented in Section 2.1 can be generalized by letting counts to have different distributions. Suppose that, conditioned on the random effects $u_i$, $y_{ij}$ has probability distribution function given by $p_j(y_{ij} | u_i)$, where $p_j(y_{ij} | u_i)$ is a Poisson, NB, ZIP or ZINB distribution. Conditioned on the random effects $u_i$, the counts $y_{ij}$ are independent, so the log-likelihood function for all data is now given by

$$l(\psi; y, u) = \sum_{i=1}^{n} \log[p(y_i; \psi)]$$

$$= \sum_{i=1}^{n} \log \left[ \prod_{j=1}^{J} p_j(y_{ij} | u_i; \psi) \right] \varphi_J(u_i; D) \, du_i. \quad (6)$$

An isolated fit of each count gives an idea of the distribution of $p_j(y_{ij} | u_i; \psi)$. Despite an increase in the number of parameters, there is no change for solving Equation (6) with respect to numerical integration compared with Equation (5).

### 2.3 Assessment of model fitting

The Akaike information criterion (AIC) is useful for comparison of regression models within a particular family of models, so for each model (Poisson, NB, ZIP and ZINB), a final model of each type was chosen using a backward criteria; the smaller the AIC value, the better is the model. One way to verify if a model is appropriate to fit the data is to use a half-normal plot. Here, we follow the approach of [37], who have suggested the use of half-normal plots as goodness-of-fit tools for models with zero-inflated count data. The plot displays the absolute values of the Pearson residuals versus half-normal scores, with simulated envelopes based on the assumed model evaluated at the parameter estimates. The Pearson residuals are defined as

$$r_{ij} = \frac{y_{ij} - \hat{E}(y_{ij})}{\hat{\text{Var}}(y_{ij})}, \quad (7)$$

where $\hat{E}(y_{ij})$ and $\hat{\text{Var}}(y_{ij})$ are the marginal mean and variances, respectively. If the model is appropriate the observed residuals fall within the simulated envelope. For the Poisson and NB models, the marginal mean is estimated as

$$\hat{E}(y_{ij}) = \exp\{x_{ij}^T \hat{\beta}_j + \hat{\sigma}_j^2/2\}, \quad (8)$$

and by $\hat{E}(y_{ij})(1 - p)$ for the ZIP and ZINB models, where $\hat{E}(y_{ij})$ is given in Equation (8).

The marginal variance of $y_{ij}$ is derived from the formula $\hat{\text{Var}}(y_{ij}) = \text{Var}(E(y_{ij} | u_i)) + E(\text{Var}(y_{ij} | u_i))$ and depends on the chosen model. If $y_{ij} | u_i \sim$ Poisson then the marginal variance
is estimated as
\[
\text{Var}(y_{ij}) = \hat{E}(y_{ij})[1 + \hat{E}(y_{ij})(\sigma^2 - 1)];
\]
if \( y_{ij} | u_i \sim NB \),
\[
\text{Var}(y_{ij}) = \hat{E}(y_{ij})[1 + \hat{E}(y_{ij})(\sigma^2(\kappa + 1) - 1)];
\]
if \( y_{ij} | u_i \sim ZIP \),
\[
\text{Var}(y_{ij}) = \hat{E}(y_{ij}) \left[ 1 + \hat{E}(y_{ij}) \left( \frac{\sigma^2}{1 - p} + 1 \right) - 1 \right];
\]
if \( y_{ij} | u_i \sim ZINB \),
\[
\text{Var}(y_{ij}) = \hat{E}(y_{ij}) \left[ 1 + \hat{E}(y_{ij}) \left( \frac{\sigma^2(k + p)}{1 - p} + 1 \right) - 1 \right].
\]

3. Application

Data were collected at the Clube Náutico Araraquara (CNA) (21°43′S, 48°01′W) located in the city of Américo Braziliense, in southeastern Brazil. Vegetation at the CNA consists of Cerrado, which is a tropical savanna formation comprising different vegetation physiognomies. Such physiognomies differ in the density and composition of plants of the woody layer and the ground layer, forming a continuum from open and dry grassland to dense forest [11,24,32]. At the CNA there are two remnants of the physiognomy locally known as ‘cerradão’, which is a dense semideciduous forest with trees 8–15 m tall and much reduced herbaceous vegetation [24,26,32]. The climate of the area has two well-defined seasons: a warm-wet season from October to March and a cool-dry season from April to September. The diet of \( G. \) microtarsus was determined by the analysis of feces sampled from individuals captured from January to December 2001 [13,16]. At the CNA, sampling was done every 15 days over two consecutive nights and individuals \( G. \) microtarsus were captured in a 7 × 7 trapping grid with 49 trapping-stations located 10 m from each other. A Sherman live-trap (dimensions 7.5 × 9.0 × 23.5 cm) was set on trees at each trapping-station about 1.75 m above ground and baited with banana and peanut butter. Individuals were marked with leg-bands and their sex was recorded. Feces on the trap floor and those defecated by individuals during manipulation were collected and preserved in 70% ethanol. Feces were transported to the laboratory and analyzed with a stereoscope. Food items detected in the feces were identified as spiders (food resource item 1), cockroaches (food resource item 2), beetles (food resource item 3), ants (food resource item 4), termites (food resource item 5) and moths (food resource item 6). Insects were identified using [5] and/or using a reference collection of insects from the study area.

Sample mean, sample variance and percentage of zeros for counts of each food item are given in

<table>
<thead>
<tr>
<th>Food resource item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean</td>
<td>0.35</td>
<td>0.39</td>
<td>1.93</td>
<td>1.93</td>
<td>10.15</td>
<td>0.52</td>
</tr>
<tr>
<td>Sample variance</td>
<td>0.63</td>
<td>0.47</td>
<td>4.24</td>
<td>4.68</td>
<td>250</td>
<td>0.70</td>
</tr>
<tr>
<td>% of zeros</td>
<td>78</td>
<td>70</td>
<td>26</td>
<td>21</td>
<td>24</td>
<td>65</td>
</tr>
</tbody>
</table>
Table 1. The sample variance is larger than the sample mean for all counts, indicating a presence of overdispersion. Food resource item 1 has the largest percentage of zeros.

Feces were sampled from a total of \( n = 37 \) (13 females and 24 males) individuals of *G. microtarsus* at the CNA field site. A total of six food resources were considered, which were labeled as 1, 2, 3, 4, 5 and 6. An exploratory analysis indicates the presence of overdispersion for all food resources, but accentuated for food resource 5, as the sample variance is much larger the sample mean. There is an indication of overdispersion due to extra zeros for food resources 1, 2 and 6 (see Table 1).

A Poisson, NB, ZIP and ZINB models with random effects have been fitted to each food resource item separately. The AIC criterion has selected the ZIP model for fitting food resource item 1, the Poisson model for fitting the food resource items 2, 3, 4 and 6 and the NB model for fitting food resource item 5. To accommodate inflation at zeros for food resource item 1, a multivariate ZIP model was fit to all data. Additionally, to accommodate the extra variability of food resource 5, a multivariate ZINB model was also adjusted to this data set.

Although it is clearly necessary to account for overdispersion due to extra zeros in these data, the point is whether it is related to all food resource counts. Guided by isolated fitting of each food resource, we further investigated two extended multivariate models according to the following structure: (1) food resource 1 follows a ZIP distribution and food resources 2, 3, 4, 5 and 6 follow a Poisson distribution; (2) the same structure of model described in item (1) with the exception that food resource 5 follows a NB distribution. The models described in items (1) and (2) have been named as ZIP + Poisson + NB and ZIP + Poisson, respectively.

Conditioned on the random effect \( u_{ij} \), the counts \( y_{ij} \) are independent, so following the notation in Section 2.1 for multivariate count models, the log-likelihood function (4) for observation \( y_{ij} \) takes \( p(y_{ij} | u_{ij}; \psi) \) as Equation (3) in a ZIP model and Equation (4) in a ZINB model, respectively. Using notation of Section 2.2 for extended multivariate count models, the log-likelihood function for \( y_{ij} \) is written as a function of

\[
\prod_{i=1}^{J} p(y_{ij} | u; \psi) = \begin{cases} \rho (1 - \rho) e^{-\theta_{01} y_{1j}} & \text{if } y_{1j} = 0 \\ e^{-\theta_{01} y_{1j}} y_{1j}^{y_{1j} - 1} e^{-\theta_{01} y_{1j}} & \text{if } y_{1j} > 0 \end{cases}
\]

for the ZIP + Poisson model and

\[
\prod_{i=1}^{J} p(y_{ij} | u; \psi) = \begin{cases} \rho (1 - \rho) e^{-\theta_{01} y_{1j}} & \text{if } y_{1j} = 0 \\ e^{-\theta_{01} y_{1j}} y_{1j}^{y_{1j} - 1} e^{-\theta_{01} y_{1j}} & \text{if } y_{1j} > 0 \end{cases}
\]

\[
\times \frac{\Gamma(y + 1/\kappa)}{y! \Gamma(1/\kappa)} (1 + \kappa \theta_{01})^{-1/\kappa} \left( 1 + \frac{1}{\theta_{01}} \right)^{-y_{1j}}
\]

for the ZIP + Poisson + NB model.

It was not possible to estimate an unrestricted covariance matrix \( D \) from these data, even with the simpler ZIP and ZINB models. As a strategy to reduce the number of variance components (here, a total of 21 components if \( D \) is unrestricted), we use common correlation, say \( \rho \), and three variance components, say \( \sigma_1^2, \sigma_2^2, \sigma_3^2 \), which represent the variability of food resource item 1, food resource items 2, 3, 4 and 6 and food resource 5, respectively. Therefore, in this application

\[
D = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 & \rho \sigma_1 \sigma_3 & \rho \sigma_1 \sigma_2 & \rho \sigma_1 \sigma_3 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_2 \sigma_1 & \sigma_2^2 & \rho \sigma_2 \sigma_3 & \rho \sigma_2 \sigma_1 & \rho \sigma_2 \sigma_3 & \rho \sigma_2 \sigma_1 \\ \rho \sigma_3 \sigma_1 & \rho \sigma_3 \sigma_2 & \sigma_3^2 & \rho \sigma_3 \sigma_1 & \rho \sigma_3 \sigma_2 & \rho \sigma_3 \sigma_1 \\ \rho \sigma_2 \sigma_1 & \rho \sigma_2 \sigma_3 & \rho \sigma_2 \sigma_1 & \sigma_2^2 & \rho \sigma_2 \sigma_3 & \rho \sigma_2 \sigma_1 \\ \rho \sigma_3 \sigma_1 & \rho \sigma_3 \sigma_2 & \rho \sigma_3 \sigma_1 & \rho \sigma_3 \sigma_2 & \sigma_3^2 & \rho \sigma_3 \sigma_1 \\ \rho \sigma_2 \sigma_1 & \rho \sigma_2 \sigma_3 & \rho \sigma_2 \sigma_1 & \rho \sigma_2 \sigma_3 & \rho \sigma_2 \sigma_1 & \sigma_2^2 \end{pmatrix}.
\]
M. Rodrigues-Motta et al.

Table 2. Estimate and SE of parameter estimates for the ZIP, ZINB and ZIP–Poisson models.

<table>
<thead>
<tr>
<th>Effect</th>
<th>ZIP</th>
<th>ZINB</th>
<th>ZIP–Poisson</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\kappa} )</td>
<td>0.114 (0.066)</td>
<td>–</td>
<td>0.496 (0.189)</td>
</tr>
<tr>
<td>( \hat{p} )</td>
<td>0.041 (0.027)</td>
<td>0.050 (0.020)</td>
<td>0.831 (0.080)</td>
</tr>
<tr>
<td>( \hat{\rho} )</td>
<td>0.868 (0.085)</td>
<td>0.843 (0.080)</td>
<td>0.897 (0.080)</td>
</tr>
<tr>
<td>( \hat{\sigma}_1 )</td>
<td>0.894 (0.385)</td>
<td>0.926 (0.382)</td>
<td>0.926 (0.382)</td>
</tr>
<tr>
<td>( \hat{\sigma}_2 )</td>
<td>0.712 (0.128)</td>
<td>0.731 (0.126)</td>
<td>0.747 (0.127)</td>
</tr>
<tr>
<td>( \hat{\sigma}_3 )</td>
<td>1.279 (0.214)</td>
<td>1.297 (0.200)</td>
<td>1.316 (0.200)</td>
</tr>
<tr>
<td>( \hat{\mu}_1 )</td>
<td>1.531 (0.457)</td>
<td>1.547 (0.458)</td>
<td>0.870 (0.622)</td>
</tr>
<tr>
<td>( \hat{\mu}_2 )</td>
<td>1.270 (0.286)</td>
<td>1.284 (0.280)</td>
<td>1.346 (0.276)</td>
</tr>
<tr>
<td>( \hat{\mu}_3 )</td>
<td>0.296 (0.188)</td>
<td>0.295 (0.182)</td>
<td>0.253 (0.179)</td>
</tr>
<tr>
<td>( \hat{\mu}_4 )</td>
<td>0.305 (0.186)</td>
<td>0.293 (0.182)</td>
<td>0.253 (0.179)</td>
</tr>
<tr>
<td>( \hat{\mu}_5 )</td>
<td>0.970 (0.186)</td>
<td>0.902 (0.180)</td>
<td>0.742 (0.273)</td>
</tr>
<tr>
<td>( \hat{\beta}_{sex} )</td>
<td>0.658 (0.209)</td>
<td>0.666 (0.202)</td>
<td>0.612 (0.204)</td>
</tr>
<tr>
<td>( \hat{\beta}_{season} )</td>
<td>-0.394 (0.133)</td>
<td>-0.457 (0.080)</td>
<td>-0.670 (0.080)</td>
</tr>
</tbody>
</table>

Table 3. Estimate and SE of the probability of zero counts of food resources 1, 2, 3, 4, 5 and 6 according to the ZINB, ZIP and ZIP + Poisson models.

<table>
<thead>
<tr>
<th>( \hat{p}_0 )</th>
<th>SE (( \hat{p}_0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINB</td>
<td>(0.74, 0.70, 0.27, 0.26, 0.18, 0.63)( ^\top )</td>
</tr>
<tr>
<td>ZIP</td>
<td>(0.74, 0.71, 0.28, 0.28, 0.22, 0.63)( ^\top )</td>
</tr>
<tr>
<td>ZIP + Poisson</td>
<td>(0.78, 0.70, 0.25, 0.25, 0.19, 0.63)( ^\top )</td>
</tr>
</tbody>
</table>

We deal with convergence problems to fit the ZIP + Poisson + NB model to this data, as different starting points produced different estimates and in some cases convergence was not achieved. The estimation of parameters of the ZIP, ZINB and ZIP–Poisson models behaved properly and convergence was achieved with small quadrature points. The best fit across the ZIP, ZINB and ZIP–Poisson models has been found with \( \beta_1 = \mu_1 \), \( \beta_2 = \mu_2 \), \( \beta_3 = \mu_3 \), \( \beta_4 = \mu_4 \), \( \beta_5 = (\mu_5, \beta_{sex}, \beta_{season})\)\( ^\top \) and \( \beta_6 = \mu_6 \). Estimates of \( \hat{\beta}_j \) and the variance components of \( \hat{D} \), along with their respective standard errors (SEs), are given in Table 2. An estimate of the probability of observing a count \( j \) being zero is \( \hat{p}_{0ij} = \hat{p} + (1 - \hat{p}) e^{-\hat{\theta}_i} \) in the ZIP model, \( (1 + \hat{\kappa}\hat{\theta})^{-1/\hat{\kappa}} \) in the ZINB model and \( \hat{p}_{0ij} = e^{-\hat{\theta}_i} \) in the Poisson model, where \( \hat{\theta}_i = \exp(x_i^\top \hat{\beta}_j + \hat{u}_{ij}) \). The \( j \)th component of the vector \( \hat{p}_0 = (\hat{p}_{0ij}, \ldots, \hat{p}_{0IJ})\)\( ^\top \) is obtained by averaging \( \hat{p}_{0ij} \) across \( i \). Results are given in Table 3.

Results in Table 2 indicate that there is an increase in the mean of consumed termites (resource 5) for being male and a decrease for dry season. The averaged estimated probabilities of observing counts equal to zero are more accurate under the ZIP–Poisson than the ZIP and ZINB models, as the predicted probability of observing a zero is closer to the percentage of observed zeros (compare Table 1 with Table 3). To further investigate the appropriateness of the ZIP, ZINB and ZIP–Poisson models for these data, simulated envelopes were constructed as suggested in Section 2.3. The graph of Figure 1 indicates there are observations falling outside the confidence bands of the simulated envelope for the ZINB model. The same happens to graph (a) of Figure 2, but in less intensity as the Pearson residuals are smaller in the ZIP model. The simulated envelope related to the ZIP + Poisson model (graph (b) of Figure 2) presents the best model fitting because (1) the range of the standard Pearson residuals is the smallest and (2) the standard Pearson residuals...
are much closer to the mean of the simulated envelope. These results select the ZIP–Poisson model as more appropriate for modeling this data set.

Selecting the ZIP–Poisson model as the best one for fitting this data, there is a chance of 78% that an animal does not eat food resource item 1 and a chance of about 50% that this animal does
not eat food resource 1 at all, either for not being exposed to it or for not foraging it. From Table 2, the correlation and the variance components are estimated as $\hat{\rho} = 0.83$, $\hat{\sigma}_1^2 = 0.81$, $\hat{\sigma}_2^2 = 0.56$ and $\hat{\sigma}_3^2 = 1.73$. The positive estimated correlation induces positive covariances in Equation (9), which from the ZIP + Poisson model is estimated as

$$D = \begin{pmatrix}
0.81 & 0.56 & 0.56 & 0.98 & 0.56 \\
0.56 & 0.46 & 0.46 & 0.82 & 0.46 \\
0.56 & 1.44 & 0.82 & 0.46 \\
0.56 & 0.82 & 0.46 \\
1.73 & 0.82 & 0.56
\end{pmatrix}.$$  

This result indicates that food resources belong to the same diet.

4. Discussion

This paper has focused on modeling multivariate correlated count data with the presence of overdispersion – including that caused by extra zeros. A multivariate count model with a common distribution across counts is initially proposed to model counts. Afterwards, the assumption of common distribution is relaxed by allowing correlated counts to follow different distributions – as it may be necessary to account for overdispersion due to extra zeros, however not for all counts. The correlation among counts is induced by using correlated random effects, with a unstructured covariance matrix. The conditional mean function depends on random effects and a set of covariates, which may be different across counts, thus adding more flexibility to the model. The modeling was applied to food resources use in a species of marsupial in a locality of the Brazilian Cerrado biome. For this data set, counts were modeled according to multivariate count models, considering the ZIP, ZINB and Poisson distributions. Simulated envelopes have selected the multivariate model described by the ZIP and Poisson distributions. In conclusion, a multivariate modeling with different distributions across counts provides a general methodology for describing count data with a wide range of dispersion properties, and thus leads to more reliable estimation of parameters of interest.

Acknowledgements

S.U. Neto and V.E. Molina have allowed access to Clube Náutico Araraquara and also have provided logical support. N.E. Oliveira and A.P.P. de Gouveia helped with the fieldwork. S.F. Reis and H. Pinheiro are partially supported by a research fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil). This research was supported by grants from FAPESP. We are thankful to the two anonymous reviewers whose comments and suggestions greatly improved the quality of the manuscript.

References

null


