Mathematical Biosciences 224 (2010) 126–130

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Bayesian semiparametric zero-inflated Poisson model for longitudinal count data

Getachew A. Dagne *

Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, 13201 Bruce B. Downs, MDC 56, Tampa, FL 33612, United States

ARTICLE INFO

Article history:
Received 27 September 2009
Received in revised form 6 January 2010
Accepted 11 January 2010
Available online 18 January 2010

Keywords:
Bayesian inference
Insecticide
P-spline
Random effects
Semiparametric model
Zero-inflation

ABSTRACT

This paper presents new methods, using a Bayesian approach, for analyzing longitudinal count data with excess zeros and nonlinear effects of continuously valued covariates. In longitudinal count data there are many problems that can make the use of a zero-inflated Poisson (ZIP) model ineffective. These problems are unobserved heterogeneity and nonlinear effects of continuously valued covariates. Our proposed semiparametric model can simultaneously handle these problems in a unified framework. The framework accounts for heterogeneity by incorporating random effects and has two components. The parametric component of the model which deals with the linear effects of time invariant covariates and the non-parametric component which gives an arbitrary smooth function to model the effect of time or time-varying covariates on the logarithm of mean count. The proposed methods are illustrated by analyzing longitudinal count data on the assessment of an efficacy of pesticides in controlling the reproduction of whitefly.

1. Introduction

Longitudinal count data with excess zeros relative to the Poisson process are common in many applications (e.g. [1–3], medical [4] and pubic health [5,6]). The zero outcomes may be attributed to either structural reasons (structural zeros) or sampling limitations (sampling zeros). For example, in the case of counts of immature whiteflies, zero counts may be recorded from plants which either never are suitable as host plants for whitefly development and reproduction (structural zeros) or for those which are suitable but no reproduction was recorded during the experimental period, thus resulting in zero-inflation. Ignoring the two types of zeros leads to model misspecification, resulting in biased parameter estimates and misleading conclusions. The most commonly used method to account for excess zeros in count data is a zero-inflated Poisson (ZIP) model [7].

Existing studies of zero-inflated models, usually ZIP, with random effects to account for correlation in repeated outcomes are generally based on restrictive assumptions such as the relationship of the conditional mean of the outcome to covariates is often assumed to be fully parametric. Specifically, the common approach is to specify linear functions of observed covariates and unobserved subject-specific effects via the log and logit link functions for the counts and zero-inflation parts [3,7,8], respectively. Although the parametric formulation enjoys simplicity, it suffers from inflexibility in modeling complex nonlinear relationship between the mean outcome and covariates. We propose to relax the linearity assumption for the regression components by using unspecified non-parametric smooth functions. For this purpose, we use penalized regression splines [9,10] for flexibly modeling zero-inflated data.

A fully Bayesian method with Markov Chain Monte Carlo (MCMC) algorithm [11,12] is used to simultaneously estimate the fixed effects parameters and non-parametric models. The extension of the parametric to semiparametric setting based on penalized regression splines enables us to estimate nonlinear effects of continuous covariates using Bayesian approach.

A key feature of this article, therefore, is that we make inference on all model components of the proposed model in a unified framework. That is, we develop a Bayesian semiparametric ZIP model that incorporates non-parametric function of nonlinear effects of continuously valued covariates, random effects, and excess zeros. Detailed description of the proposed model is presented in Section 2. Section 3 illustrates an application of the proposed methods to assess the efficacy of pesticide in reducing the reproduction of whitefly over time. Section 4 ends the article with a conclusion.

2. Bayesian semiparametric zero-inflated Poisson model

2.1. The basic model specification

For specifying the proposed models, let the discrete response variable y_{ijk} be the count for the ith subject in the jth ($j = 1, 2, \ldots J$) block during time period k ($k = 1, 2, \ldots T_{ij}$). In our application (see Section

* Tel.: +1 813 974 6680; fax: +1 813 974 4719.
E-mail address: gdagne@health.usf.edu

0025-5564/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2010.01.004

3) on reproduction of whitefly after treatment application, we let n_j denote the number of plants in block j during week t so that $\sum_{j=1}^{J} n_j$ gives the total number plants and $\sum_{j=1}^{J}\sum_{i=1}^{n_j} T_{ij}$ denoting the total number of observations. We assume that for each observed count, y_{ijk}, there is a latent random variable D_{ijk} with the observed binary random variable $D_{ijk} = I(D_{ijk} > 0)$ representing either the zero-state or the Poisson-state from which each observation is drawn. The D_{ijk} are assumed independently drawn from a Bernoulli distribution with parameter p_{ijk}, such that $\Pr(D_{ijk} = 1) = p_{ijk}$ if $y_{ijk} = 0$ comes from the zero-state corresponding to never suitable for reproduction, and $\Pr(D_{ijk} = 0) = 1 - p_{ijk}$ if y_{ijk} is generated from the Poisson-state. Given the values of D_{ijk}, covariate vectors and unobserved effects, y_{ijk} is distributed as zero-inflated Poisson distribution, $f(y_{ijk})$, given by

$$y_{ijk} \sim \begin{cases} 0, & \text{with probability } p_{ijk} \\ \text{Poisson}(\lambda_{ijk}), & \text{with probability } 1 - p_{ijk}, \end{cases} \tag{1}$$

where $\text{Poisson}(\lambda_{ijk})$ is defined as $exp(-\lambda_{ijk})\lambda_{ijk}^{y_{ijk}}/y_{ijk}!$.

The parameters λ and p in (1), which correspond to the Poisson and inflation components, can be modeled as functions of covariates and unobserved effects via the log and logit links as follows:

$$\log(\lambda_{ijk}) = \mathbf{x}'_{ijk}\beta + g(t_{ijk}) + u_i, \tag{2}$$

and

$$logit(p_{ijk}) = \mathbf{z}'_{ijk}\gamma + h(w_{ijk}) + v_i. \tag{3}$$

In the models (2) and (3), $\mathbf{x}_{ijk}$ and $\mathbf{z}_{ijk}$ are vectors of covariates which are not necessarily the same for the zero-state and the Poisson-state, respectively, and $h(\cdot)$ and $g(\cdot)$ represent unknown smooth nonlinear functions of observable continuous covariates t_{ijk} and w_{ijk}. Here β and γ are the associated vectors of unknown coefficient parameters, and (u_i, v_i) are the subject level unobserved effects which are assumed to be independent and normally distributed with mean 0 and variances σ_u^2 and σ_v^2, respectively.

The likelihood function for the mixed zero-inflated model, conditional on all covariates ($\mathbf{x}^\dagger$) and unobservables ($\mathbf{u}^\dagger$), takes the general form:

$$\begin{aligned} \mathscr{L}(y, D|\mathbf{x}^\dagger, \mathbf{u}^\dagger, \Psi) = \prod_{ijk} &[\Pr(D_{ijk} = 1) + \Pr(D_{ijk} = 0) \\ &\times \Pr(y_{ijk} = 0)]^{D_{ijk}} [\Pr(D_{ijk} = 0)(f(y_{ijk})]^{1-D_{ijk}}, \end{aligned} \tag{4}$$

where Ψ is a vector of parameters, and the distribution of the count variable y_{ijk}, conditional on covariates and unobserved effects, is assumed to be Poisson.

2.2. Bayesian semiparametric ZIP mixed-effects models

In this sub-section, we present the modeling of the unknown non-parametric functions $h(\cdot)$ and $g(\cdot)$ in (2) and (3). We use penalized spline smooth functions to approximate the non-parametric functions [13–15]. Penalized spline fitting as smoothing technique has become very popular recently because of the link between smoothing functions and linear mixed models which makes the procedure so attractive. Based on a suggestion by Crainiceanu et al. [9], this connection is made by using cubic spline basis for representing $g(t_{ij})$ by $\alpha_0 + \alpha_1 t_{ij} + \sum_{s=1}^{K} b_s|t_{ij} - \kappa_s|^3$, where $\kappa_1 < \kappa_2 < \cdots, \kappa_K$ are fixed knots, and these knots are typically placed at quantiles of the distribution of unique values of the covariate t_{ij}. With respect to the dimension K we follow Ruppert et al. [16] recommendation that the actual choice of K and the location of knots have little influence on the resulting penalized fit as long as K is large. The value of K is chosen between 5 and 35 to ensure enough flexibility [16]. The random coefficient of the P-spline functions b is assumed to be normally distributed with mean zero and variance σ_b^2. The smooth function for $h(\cdot)$ is also defined in a similar fashion with different parameters and knots. For estimation, we use a Bayesian approach. A Bayesian penalized splines has the advantage of allowing for simultaneous estimation of smooth functions and smoothing parameters.

2.2.1. Prior distributions

Under Bayesian framework, we need to specify prior distributions for unknown parameters in the models (2) and (3) as follows.

For fixed effects β and γ, we choose weakly informative normal priors. That is, $\beta \sim N(0, \Sigma_\beta)$ and $\gamma \sim N(0, \Sigma_\gamma)$, where Σ_β and Σ_γ are assessed to have large values to make the prior distributions weakly informative but proper. We also assume inverse gamma (IG) priors for the variances of the random effects v and u. That is, $\sigma_u^2 \sim IG(.01, .01)$ and $\sigma_v^2 \sim IG(.01, .01)$ so that the gamma distribution has mean 1 and variance 100. Likewise, both non-parametric functions of (2) and (3) have unknown parameters for fixed effects components and random effects components. For the fixed components, α_{0q} and α_{1q} jointly follow a multivariate normal distribution with means 0 and variance Σ_{α_q} for $q = 1$ for $g(\cdot)$ and $q = 2$ for $h(\cdot)$. For the variances of the P-spline random coefficients, $\sigma_g^2 \sim IG(.01, .01)$ and $\sigma_h^2 \sim IG(.01, .01)$.

Given that the prior distributions for parameters have been assessed, the next procedure is combine the likelihood function in (4) with priors to make a Bayesian inference. This procedure is implemented using MCMC algorithm.

2.2.2. Implementation of MCMC algorithm

The MCMC simulation sampling was implemented using WinBUGS software [17], and the program codes are available from the author upon request. When the MCMC implementation is applied to the pesticide data (see Section 3), convergence of the MCMC samples is assessed using standard tools within WinBUGS software (trace plots, ACF plots, as well as Gelman–Rubin convergence diagnostic). After an initial 10000 burn-in iterations, 10000 samples with thinning 20 are obtained to make inference. After fitting these models, we also use a Bayesian model selection technique to choose the best model that fits the data in Section 3.

2.3. Model selection

We use a Bayesian model selection procedure to choose a model that fits the data well. One such a procedure is the deviance information criterion (DIC), suggested by Spiegelhalter et al. [18], which generalizes the Akaike information criterion (AIC) to a Bayesian approach. Like AIC, DIC trades off a measure of model adequacy against a measure of complexity, and also is easy to calculate and apply to a wide range of statistical models. It is based on the posterior distribution of the log-likelihood in the Bayesian framework. To compare between various models (e.g., models with parametric and non-parametric time effects) we use DIC. There are many other Bayesian approaches to model selection (e.g., posterior model probabilities, Bayes factor [19], posterior predictive checks [20]). However, some of these methods are not well defined with vague prior while the others are not automatic nor easily reduced to a unique, single number summary [21]. In addition, although hierarchical Bayesian methods implemented via MCMC procedures enable the fitting of such models, a formal comparison of their fit is hampered by their large size and often improper specifications. By using a complexity measure for the effective number of parameters that is based on an information theoretic argument, DIC avoids such problems. We will use the recently developed DIC [18] for model comparison in this paper. Guo and Carlin [21] gave several advantages for choosing DIC as model selection criteria.

DIC is given as $DIC = goodness\text{-}of\text{-}fit + penalty\ for\ complexity$ where the "goodness-of-fit" is measured by the deviance for Ψ,