Spatial and Spatio-temporal Epidemiology 2 (2011) 23–33

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Modeling type 1 and type 2 diabetes mellitus incidence in youth: An application of Bayesian hierarchical regression for sparse small area data

Hae-Ryoung Song [a], Andrew Lawson [b], Ralph B. D'Agostino Jr. [c], Angela D. Liese [a,*]

[a] *Department of Epidemiology and Biostatistics and Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA*

[b] *Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC, USA*

[c] *Department of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, USA*

ARTICLE INFO

Article history:
Received 23 December 2009
Revised 27 August 2010
Accepted 1 September 2010

Keywords:
Sparse data
Diabetes
Bayesian
Sparse Poisson convolution model
Sparse Poisson MCAR model

ABSTRACT

Sparse count data violate assumptions of traditional Poisson models due to the excessive amount of zeros, and modeling sparse data becomes challenging. However, since aggregation to reduce sparseness may result in biased estimates of risk, solutions need to be found at the level of disaggregated data. We investigated different statistical approaches within a Bayesian hierarchical framework for modeling sparse data without aggregation of data. We compared our proposed models with the traditional Poisson model and the zero-inflated model based on simulated data. We applied statistical models to type 1 and type 2 diabetes in youth 10–19 years known as rare diseases, and compared models using the inference results and various model diagnostic tools. We showed that one of the models we proposed, a sparse Poisson convolution model, performed better than other models in the simulation and application based on the deviance information criterion (DIC) and the mean squared prediction error.

1. Introduction

Sparse data is often encountered in epidemiologic studies. For instance, even though diabetes mellitus ranks as the third most common chronic disease among youths, it is still a very rare disease with an estimated incidence of 24.3 per 100,000 person-years at risk (Dabelea et al., 2007). One common solution for modeling sparse data is to aggregate data to a larger spatial or temporal unit, and then model the aggregated data. Previous studies of diabetes mellitus in youth have frequently employed temporal aggregation, basing spatial analyses on 10 or more years of incidence data (Patterson and Waugh, 1992; Samuelsson et al., 2004; Feltbower et al., 2003; Staines et al., 1997; Schober et al., 2003; Waldhor et al., 2003; Rytkonen et al., 2003). While in the past most spatial analyses of diabetes incidence relied on describing incidence rates by region (Patterson and Waugh, 1992; Samuelsson et al., 2004; Feltbower et al., 2003), more recently standard Poisson convolution models have been applied in a Bayesian framework (Schober et al., 2003; Waldhor et al., 2003; Cardwell et al., 2006). However, it is well known that spatial or temporal aggregation of data often causes ecological bias (Piantadosi et al., 1988; Greenland and Robins, 1994; Wakefield, 2004; Lawson, 2006). The ecological fallacy (Firebaugh, 2001; Freedman, 2001) occurs when inferences are made about individual level associations based on

Abbreviations: ZIP, zero-inflated Poisson; SC, South Carolina; DIC, deviance information criterion; MSPE, mean squared prediction error; EZIP, extended ZIP; SPC, sparse Poission convolution; CAR, conditional autoregressive; MCAR, multivariate CAR; SPMCAR, sparse Poisson MCAR; RR, relative risk.

* Corresponding author. Address: Department of Epidemiology and Biostatistics and Center for Research in Nutrition and Health Disparities, Arnold School of Public Health, University of South Carolina, 800 Sumter Street, Columbia, SC 29208, USA. Tel.: +1 803 777 9414; fax: +1 803 777 2524.

E-mail address: liese@sc.edu (A.D. Liese).

1877-5845/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.sste.2010.09.008

aggregate data. Another effect is that aggregation increases the spatial correlation between observation units. Thus, there is a need to identify methods suitable for dealing with sparse data at lower levels of aggregation.

The log linear Poisson model which is typically applied to aggregated count data (Lawson, 2006; Besag and Kooperberg, 1995) cannot, however, be applied successfully to sparse or disaggregated data. In sparse data, overdispersion is common when the variance is larger than the mean due to the excess amount of zeros.

In the widely used zero-inflated Poisson (ZIP) model (Cheung, 2002; Martin et al., 2005; Lawson, 2008), zero observed counts are divided into excess zero counts and nonexcess zero counts. Excess zero counts are regarded as zero counts which are observed in the process excessively and cannot be modeled by the Poisson distribution; while nonexcess zero counts are zero counts which are derived from a Poisson distribution. While the traditional ZIP models do not model the excess zero counts based on the Poisson model, we may suggest a more sophisticated approach which models the excess zero counts as well as nonexcess zero counts using a special form of the Poisson model for better inference.

We investigate different statistical approaches within a Bayesian hierarchical framework for modeling sparse data without aggregation of data. We modify the ZIP model, and suggest several new statistical models. The Bayesian approach is regarded as a flexible modeling approach compared to the frequentist approach in that it combines both data information and prior information in inference through prior distributions of each parameter, and it enables the building of a complex model by using a hierarchical structure. We compare our proposed models with the traditional Poisson convolution model and the ZIP model through simulation studies, and apply our statistical models to data on the incidence of type 1 and type 2 diabetes mellitus in youth aged 10–19 years in South Carolina (SC) as part of a project on the spatial epidemiology of diabetes in youth (Liese et al., 2010). Model evaluation is conducted based on model diagnostic tools such as the deviance information criterion (DIC) and the mean squared prediction error (MSPE) (Lawson, 2008; Banerjee et al., 2006).

2. Statistical models

Observed counts are typically fitted by the Poisson model with a log link which establishes a log linear relationship between the mean of the Poisson model and covariates.

In the log linear Poisson model, we include log expected counts as an offset to model the relative risk of disease, and also add other covariates to capture confounding effects and random effects to explain the additional variation that cannot be captured by covariates. Let $y_i = 1,\ldots,n$ be the observed count of the ith region. The standard Poisson regression model for modeling count data is defined as:

$$y_i \sim Pois(\lambda_i) = \frac{exp(-\lambda_i)\lambda_i^{y_i}}{Y_i!}$$
$$\log(\lambda_i) = \log(E_i) + \alpha + X_i\beta + u_i + v_i$$

where E_i is an expected count, α is an intercept, X_i is a matrix of covariates, β is a vector of parameters associated with individual covariates, and u_i and v_i are spatially correlated and uncorrelated random effects, respectively. This is the classic convolution model with covariates first proposed by Besag et al. (1991).

In the presence of an excess amount of zeros in data, the Poisson model is not a suitable model to apply to data because the key model assumption of the equality of the mean and the variance of the Poisson model is not met. The ZIP model is a frequently suggested model, where a mixture model of a proportion $1 - p$ of excess zeros, and a proportion p of nonexcess zero and nonzero counts is assumed. Excess zero counts are zero counts that are not derived from a Poisson distribution, and nonexcess zero counts are natural zero counts which are derived from a Poisson distribution. In the ZIP model, the Poisson model is fitted by utilizing only a proportion of nonexcess zeros and total nonzero counts. The ZIP model is defined as:

$$\begin{aligned}
&y_i \sim Pois(\lambda_i)\\
&\lambda_i = I_i * \mu_i\\
&\log(\mu_i) = \log(E_i) + \alpha + X_i\beta + u_i + v_i\\
&I_i \sim Bernoulli(p)\\
&P \sim beta(1,1)
\end{aligned}$$

where I is the indicator to distinguish excess zero counts and nonexcess zero or nonzero counts ($I = 0$ for excess zero counts, and $I = 1$ for nonexcess zero or nonzero counts) and p is the probability of nonexcess zero counts. A review of zip models can be found in Ghosh et al. (2006).

2.1. Novel models

Here, we propose several models which use the information from expected counts in modeling the proportion of excess zeros, and refer to them as extended ZIP (EZIP) models. In EZIP1, we model the probability of excess zero counts $1 - p_i$ as a function of expected counts E_i: $p_i = E_i/(\delta + E_i)$, where δ is a threshold to distinguish excess zeros due to the low values of expected counts from nonexcess zero counts. If we observe an expected count in a region which is larger than δ, p_i becomes close to 1 which indicates that the probability of observing an excess zero count in a region becomes small. On the other hand, for an expected count smaller than δ, we have more probability of observing excess zero counts. The threshold can be chosen to a fixed particular value or it can be estimated within the model. The EZIP1 is defined as:

$$\begin{aligned}
&y_i \sim Pois(\lambda_i)\\
&\lambda_i = I_i * \mu_i\\
&\log(\mu_i) = \log(E_i) + \alpha + X_i\beta + u_i + v_i\\
&I_i \sim Bernoulli(p)\\
&P \sim E_i/(\delta + E_i)
\end{aligned}$$

In addition to the EZIP1 model we also examined two other models that are variants of this model: EZIP2 and EZIP3. The EZIP2 model adopts the information on expected counts directly in modeling the Poisson distribution as well as the proportion of excess zero counts. While the