COACTION PUBLISHING

Analysis of zero-inflated and overdispersed time series: an application to syphilis surveillance in the United States

Ming Yang[1]*, Joseph Cavanaugh[1] and Philip Polgreen[2,3]

[1]Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, IA, USA; [2]Division of Infectious Diseases, Department of Internal Medicine, Carver College of Medicine, University of Iowa, Iowa City, IA, USA; [3]Department of Epidemiology, College of Public Health, University of Iowa, Iowa City, IA, USA

Objective

The purpose of this study is to develop novel statistical methods to analyze zero-inflated and overdispersed time series consisting of count data.

Introduction

Time series data involving counts are frequently encountered in many biomedical and public health applications. For example, in disease surveillance, the occurrence of rare infections over time is often monitored by public health officials, and the time series data collected can be used for the purpose of monitoring changes in disease activity. For rare diseases with low infection rates, the observed counts typically contain a high frequency of zeros (zero-inflated), but the counts can also be very large (overdispersed) during an outbreak period (1). Failure to account for zero-inflation and overdispersion in the data may result in misleading inference and the detection of spurious associations.

Methods

Under the partial likelihood framework (2), we develop a class of regression models for zero-inflated and overdispersed count time series based on the conditional zero-inflated negative binomial (ZINB) distribution with probability mass function defined as follows:

$$f(y_t; k,\ \mu_t, \pi_t | \mathcal{F}_{t-1}) = \pi_t I(y_t = 0) + (1-\pi_t)\frac{\Gamma(k+y_t)}{\Gamma(k) y_t!}\left(\frac{k}{k+\mu_t}\right)^k \left(\frac{\mu_t}{k+\mu_t}\right)^{y_t}$$

The ZINB distribution is very general; it is a two-component mixture of the NB distribution with a point mass at zero. It reduces to the NB distribution when the zero-inflation parameter is zero and the zero-inflated Poisson (ZIP) distribution as the dispersion parameter goes to infinity.

Results

We applied the methodologies proposed above to monthly syphilis data in the United States from 1995 to 2009 (http://www.cdc.gov/mmwr/). During the study period, a high proportion of zeros and some large positive counts were observed in most of the 66 surveillance locations. Among the four candidate distributions (Poisson, NB, ZIP and ZINB), we find that the ZINB distribution is most frequently favored in terms of Akaike's information criterion (AIC) (3). In contrast, we find that the Poisson distribution is never selected for any of the surveillance locations (Table 1).

Conclusions

Although the Poisson distribution has been used widely in public health practice, its performance becomes unreliable in the presence of zero-inflation and overdispersion. The ZINB distribution is an attractive alternative to the Poisson distribution, as it provides a unified approach to model zero-inflated and overdispersed count time series in a variety of disciplines.

Table 1. Model selection results for the 66 surveillance locations

Distribution	*Poisson*	*Negative binomial*	*Zero-inflated Poisson*	*Zero-inflated negative binomial*
Frequency	0	11	6	49

Keywords

Syphilis surveillance: time series; zero-inflation

References

1. Yau KKW, Lee AH, Carrivick PJW. Modeling zero-inflated count series with application to occupational health. Comput Meth Progr Biomed. 2004;74:47–52.
2. Kedem B, Fokianos K. Regression models for time series analysis. New Jersey: Wiley, 2002.
3. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19:716–23.

*Ming Yang
E-mail: ming-yang@uiowa.edu

Citation: Emerging Health Threats Journal 2011, **4**: 11117 - DOI: 10.3402/ehtj.v4i0.11117