Computational Statistics and Data Analysis 61 (2013) 67–82

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Score tests for zero-inflation and overdispersion in two-level count data

Hwa Kyung Lim [a], Juwon Song [b,*], Byoung Cheol Jung [c]

[a] Institute of Statistics, Korea University, Seoul, Republic of Korea
[b] Department of Statistics, Korea University, Seoul, Republic of Korea
[c] Department of Statistics, University of Seoul, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 7 April 2011
Received in revised form 10 May 2012
Accepted 7 November 2012
Available online 17 November 2012

Keywords:
Zero-inflation
Overdispersion
Generalized linear mixed models
Zero-inflated negative binomial
Score test
Bootstrap

ABSTRACT

In a Poisson regression model in which observations are either clustered or represented by repeated measurements of counts, the number of observed zero counts is sometimes greater than the expected frequency by the Poisson distribution and overdispersion may remain even after modeling excess zeros. The zero-inflated negative binomial (ZINB) mixed regression model is suggested to analyze such data. Previous studies have proposed score statistics for testing zero-inflation and overdispersion separately in correlated count data. Here, we also deal with simultaneous score tests for zero-inflation and overdispersion in two-level count data by using the ZINB mixed regression model. Score tests are suggested for (1) zero-inflation in the presence of overdispersion, (2) overdispersion in the presence of zero-inflation, and (3) zero-inflation and overdispersion simultaneously. The level and power of score test statistics are evaluated by a simulation study. The simulation results indicate that score test statistics may occasionally underestimate or overestimate the nominal significance level due to variation in random effects. This study proposes a parametric bootstrap method to overcome this problem. The simulation results of the bootstrap test indicate that score tests hold the nominal level and provide good power.

1. Introduction

In fields such as medicine, public health, epidemiology, sociology, psychology, engineering, and agriculture, among others, the analysis of count data is a topic of major interest. For count data, Poisson regression models have been widely used to explain the relationship between the outcome variable of interest and a set of explanatory variables. However, there are some cases in which the number of observed zero counts exceeds the expected frequency by the Poisson distribution. In such cases, a standard Poisson model may not perform well. A fair number of statistical methods have been developed to address count data with extra zeros. Böhning (1998) reviews the related literature and presented some examples from a wide variety of disciplines. A popular approach for analyzing count data with excess zeros is to use the zero-inflated Poisson (ZIP) regression model by Lambert (1992). The ZIP regression model is a mixture of the Poisson distribution and a degenerate component of the point mass at zero. Van den Broek (1995) proposes a score test for zero-inflation under a Poisson distribution. Deng and Paul (2000) and Jansakul and Hinde (2002) extend this to the ZIP regression model with covariates, and Deng and Paul (2000) further consider the zero-inflated binomial (ZIB) regression model with covariates.

* Corresponding author. Tel.: +82 2 3290 2241; fax: +82 2 3290 2241.
E-mail addresses: hklim@korea.ac.kr (H.K. Lim), jsong@korea.ac.kr (J. Song), bcjung@uos.ac.kr (B.C. Jung).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.11.006

Both zero-inflation and dependency can often be present in hierarchical count data in which observations are either clustered or repeatedly measured from individual subjects. For example, subjects sampled from a common habitat (called a cluster) such as families, schools, and communities are more likely to be similar to one another than those sampled across different habitats, resulting in correlated responses within the cluster. Dependency among responses can be explained by hierarchical structures through the use of random effects. Hall (2000), Yau and Lee (2001), Hur et al. (2002), and Wang et al. (2002) consider ZIP regression models with cluster-specific random effects to address the heterogeneous variances among clusters. Xiang et al. (2006) propose a score test for zero-inflation in correlated count data, and Lee et al. (2006) extend the ZIP regression model to a multilevel ZIP regression model with random effects. Recently, Moghimbeigi et al. (2009) propose a score test for zero-inflation in multilevel count data.

In Poisson data with too many zeros, the variance often exceeds the mean, causing overdispersion. Although the ZIP regression model can handle excess zeros for Poisson data, overdispersion may remain even after modeling excess zeros, and consequently ZIP parameter estimates can be severely biased. In such a case, the use of a zero-inflated negative binomial (ZINB) distribution can be a good alternative. Ridout et al. (2001) consider overdispersion in count data and propose a score test for testing the ZIP regression model against ZINB alternatives. For hierarchical or correlated count data, it is especially true that ZIP parameter estimates can be severely biased when nonzero counts are overdispersed. Xiang et al. (2007) propose a score test for assessing overdispersion based on the ZINB mixed model, while those of Xie et al. (2009) and Yang et al. (2010) focus on the zero-inflated generalized Poisson (ZIGP) mixed model. However, a simultaneous score test for zero-inflation and overdispersion in the ZINB mixed model or the ZIGP mixed model has not been proposed. Deng and Paul (2005) consider simultaneous score tests for zero-inflation and overdispersion in the ZINB regression model, but their model does not involve random effects for clustered count data.

In this paper, we deal with score tests for zero-inflation and/or overdispersion in two-level count data fitted by the ZINB mixed regression model. We propose score tests for (1) zero-inflation in the presence of overdispersion, (2) overdispersion in the presence of zero-inflation, and (3) zero-inflation and overdispersion simultaneously. Section 2 describes the ZINB mixed regression model. Section 3 suggests score tests for zero-inflation and/or overdispersion in the ZINB mixed regression model. The results of a simulation study checking the adequacy of the χ^2 distribution as an asymptotic distribution of the score test statistic indicate that the asymptotic χ^2 distribution performs poorly for large variation in random effects and small samples. To overcome this difficulty, Section 4 proposes a parametric bootstrap method. Section 5 compares the level and power of score test statistics by using the χ^2 approximation method and a bootstrap method through a simulation study. To illustrate the model selection procedure based on the proposed tests, Section 6 presents an analysis of the DMFT index, and Section 7 concludes by discussing the findings and suggesting some interesting avenues for future research.

2. ZINB mixed regression model

Let Y_{ij} be the jth response of a count variable from the ith cluster. Then the ZINB distribution can be written as

$$P(Y_{ij}=y_{ij})=\begin{cases}\phi_{ij}+(1-\phi_{ij})(1+\alpha\lambda_{ij})^{-1/\alpha} & \text{if } y_{ij}=0\\ (1-\phi_{ij})\dfrac{\Gamma\left(y_{ij}+1/\alpha\right)}{y_{ij}!\;\Gamma\left(1/\alpha\right)}(1+\alpha\lambda_{ij})^{-1/\alpha}\left(1+\dfrac{1}{\alpha\lambda_{ij}}\right)^{-y_{ij}} & \text{if } y_{ij}>0\end{cases}$$

for $i=1,\ldots,m$ and $j=1,\ldots,n_i$, where m is the number of clusters, n_i is the number of observations for cluster i, the total number of responses is $N=\sum_{i=1}^{m}n_i$ and $\alpha>0$ is an overdispersion parameter. Let $0<\phi_{ij}<1$ such that the ZINB distribution allows for more zeros than the negative binomial ($\phi_{ij}=0$), while $\phi_{ij}<0$ corresponds to the zero-deflated situation (Dietz and Böhning, 2000). When λ_{ij} indicates the mean of the Poisson distribution, the mean and variance of the ZINB response variable are given by

$$E(Y_{ij})=\left(1-\phi_{ij}\right)\lambda_{ij},$$

$$Var(Y_{ij})=(1-\phi_{ij})\lambda_{ij}(1+\phi_{ij}\lambda_{ij}+\alpha\lambda_{ij}).$$

The parameters ϕ_{ij} and λ_{ij} can be modeled by linking linear predictors as follows:

$$g(\phi_{ij})=\xi_{ij}=w_{ij}'\gamma+u_i,$$

$$h(\lambda_{ij})=\eta_{ij}=x_{ij}'\beta+v_i,$$

where w_{ij} and x_{ij} are explanatory variables, and γ and β are the corresponding $p\times1$ and $q\times1$ vectors of regression coefficients for ϕ_{ij} and λ_{ij}, respectively. The same explanatory variables for w_{ij} and x_{ij} can be used, and then more parsimonious models, such as that in Lambert (1992), would be possible to be developed. In these models, responses in different clusters are assumed to be independent, whereas those within the same cluster are likely to be correlated. To accommodate inherent correlations within clusters, random effects u_i and v_i are incorporated into the linear predictor ξ_{ij} for the zero-inflation model and into η_{ij} for the Poisson model, respectively. The random effects $u=(u_1,\ldots,u_m)'$ and $v=(v_1,\ldots,v_m)'$ are assumed to be independently distributed as $N(0,\sigma_u^2)$ and $N(0,\sigma_v^2)$, respectively, and $g(\cdot)$ and $h(\cdot)$ are known as link functions. Since