Computational Statistics & Data Analysis 40 (2002) 75–96
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Score Tests for Zero-Inflated Poisson Models

N. Jansakul[a,*], J.P. Hinde[b]

[a] *Department of Mathematics, Prince of Songkla University, Hatyai, Songkla 90112, Thailand*
[b] *School of Mathematical Sciences, Laver Building, Exeter University, North Park Road, Exeter EX4 4QE, UK*

Received 1 January 2001; received in revised form 1 November 2001; accepted 1 November 2001

Abstract

In many situations count data have a large proportion of zeros and the zero-inflated Poisson regression (ZIP) model may be appropriate. A simple score test for zero-inflation, comparing the ZIP model with a constant proportion of excess zeros to a standard Poisson regression model, was given by van den Broek (Biometrics, 51 (1995) 738–743). We extend this test to the more general situation where the zero probability is allowed to depend on covariates. The performance of this test is evaluated using a simulation study. To identify potentially important covariates in the zero-inflation model a composite test is proposed. The use of the general score test and the composite procedure is illustrated on two examples from the literature. The composite score test is found to suggest appropriate models.

Keywords: Count data; Score test; Poisson regression; Zero-inflation; Composite score test

1. Introduction

Poisson regression provides a standard model for the analysis of count data. However, in practice, count data are frequently more variable than specified by the Poisson model and are described as overdispersed. There are various mechanisms that can lead to overdispersion and a number of different overdispersed count regression models have been proposed; see Hinde and Demétrio (1998). This paper concentrates on situations where the number of observed zero counts is larger than expected under a standard Poisson model, giving a very specific type of overdispersion. A simple mixture model, the zero-inflated Poisson (ZIP) model, can be used for data of this form—this is a

* Corresponding author. Fax: +66-074-212921.
E-mail addresses: jnaratip@ratree.psu.ac.th (N. Jansakul), j.p.hinde@exeter.ac.uk (J.P. Hinde).

0167-9473/02/$ - see front matter
PII: S0167-9473(01)00104-9

combination of a Poisson model and a degenerate distribution at zero. The ZIP model can incorporate explanatory variables in both the zero process and the Poisson model.

This paper considers the use of score tests to compare the Poisson model fits to ZIP models, providing a simple check for zero-inflation. A test of this form was first proposed by van den Broek (1995) for the special case of a ZIP model with a constant proportion of excess zeros. Here we extend this test to the general situation where the zero probability is allowed to depend on covariates. To screen potential covariates we also propose the use of a composite test procedure based on score tests against a collection of models for the zero process. This is seen to give a simple and effective way of selecting a model for the zero process.

The layout of the paper is as follows. In Section 2 we present a short review of Poisson and ZIP regression models. In Section 3 we consider the general problem of testing for zero-inflation and derive a new score test extending the result of van den Broek (1995). The performance of this test is considered through a small simulation study. A particular advantage of the score test is that it only requires parameter estimates from the Poisson model; there is no need to fit the more complex ZIP model. This aspect is particularly useful as it allows a large range of models for the zero process to be tested from a single fit for the Poisson model. This idea is used in Section 4 to derive composite tests of models for the proportion of excess zeros. Section 5 illustrates the use of these tests on some real datasets and shows that the composite tests are useful for suggesting appropriate models. The paper ends with a discussion of extensions and other contexts in which the ideas could be implemented.

2. Poisson and ZIP regression models

Suppose that $Y_i,\ i=1,\dots,n$, are counts with means μ_i, and that $\mathbf{x}_i=(x_{i1},x_{i2},\dots,x_{ip})^{\mathrm{T}}$ is an associated vector of covariates, with x_{i1} typically equal 1 to include the usual constant term in the model. The standard Poisson regression model assumes that $Y_i \sim \mathrm{Pois}(\mu_i)$, with variance function

$$\mathrm{Var}(\mu_i) = \mu_i$$

and μ_i is modelled through the canonical log link function by

$$\eta_i = \log(\mu_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. This is a standard generalized linear model and the maximum likelihood estimate of $\boldsymbol{\beta}$ is easily obtained using iteratively reweighted least squares (IRLS), see McCullagh and Nelder (1989).

For an appropriate well-fitting model, we would expect

$$\text{Residual deviance} \approx \text{Residual degrees of freedom}.$$

If the residual deviance exceeds its degrees of freedom, the observed variance may be greater than the nominal Poisson variance. That is, the data could be overdispersed. If the residual deviance is less than its degrees of freedom, it implies that there is underdispersion in the count data, i.e. the actual variance is less than the nominal