Computational Statistics and Data Analysis 61 (2013) 99–106

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Conditional tests for homogeneity of zero-inflated Poisson and Poisson-hurdle distributions

Edward J. Bedrick [a,*], Anwar Hossain [b]

[a] Department of Internal Medicine, University of New Mexico, Albuquerque, NM 87131, United States

[b] Department of Mathematics, New Mexico Institute of Mining and Technology, Socorro, NM 87801, United States

ARTICLE INFO

Article history:
Received 14 February 2011
Received in revised form 8 November 2012
Accepted 8 November 2012
Available online 16 December 2012

Keywords:
Contingency table
Overdispersion
Stirling number
Truncated Poisson distribution

ABSTRACT

We develop two conditional tests for homogeneity of zero-inflated Poisson (ZIP) and Poisson-hurdle distributions. A Monte Carlo method is proposed for approximating the reference distributions of these tests. The techniques are applied to two examples.

1. Introduction

There has been considerable recent interest in modeling count data that has an excess number of zeros. For example, the zero-inflated Poisson (ZIP) model assumes that the counts arise from a mixture model where one component in the mixture is a point mass at zero and the other component follows a Poisson distribution. The mixing of the distributions leads to an excess number of zero counts relative to a Poisson model. The ZIP model has found application in diverse disciplines such as manufacturing (Lambert, 1992), ecology (Welsh et al., 1996), economics (Cameron and Trivedi, 1998), and dental epidemiology (Böhning et al., 1999). The standard frequentist approach to inference for the ZIP model is based on large sample maximum likelihood theory. Alternative models such as the zero-inflated negative binomial model (Ridout et al., 2001) have been proposed but they are not as widely used as the ZIP model.

The widespread availability of powerful personal computers has led to the development and routine use of exact methods for analyzing discrete data. Standard statistical packages such as SAS (2007) and `Stata` (StataCorp., 2007) have exact procedures while two software packages, `StatXact` (Cytel Inc., 2007a) and `LogXact` (Cytel Inc., 2007b), emphasize exact methods. Although exact methods partly eliminate the need for large sample approximations, there is an ongoing debate about the relative merits of conditional versus unconditional approaches to exact inference (Mehta and Hilton, 1993). Furthermore, existing exact methods are primarily limited to binomial, multinomial, Poisson, or clustered binary data (Corcoran et al., 2001; Han et al., 2004).

We show that exact conditional inference is possible for the ZIP model and the closely related Poisson-hurdle model. We focus attention on the problem of testing homogeneity of distributions, but suggest in the concluding section how the methods might be extended to more general regression models. We show that the reference distribution for testing

* Correspondence to: Division of Epidemiology and Biostatistics, Department of Internal Medicine, MSC10 5550, University of New Mexico, Albuquerque, NM 87131, United States. Tel.: +1 011 505 272 2520.
E-mail addresses: EBedrick@salud.unm.edu (E.J. Bedrick), hossain@nmt.edu (A. Hossain).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.11.009

homogeneity can be factored into two components. The first component is the reference distribution for testing the hypothesis that the probability of a zero count is constant across populations. The second piece is the reference distribution for testing homogeneity of truncated Poisson distributions. In related work, Tse et al. (2009) derive large sample likelihood ratio tests for the homogeneity of two ZIP distributions. Our approach allows for an arbitrary number of distributions and does not rely on asymptotic approximations.

The remainder of the paper is organized as follows. Section 2 provides background on the ZIP and Poisson-hurdle models and derives the form of two conditional tests for homogeneity of distributions. Section 3 develops a method for simulating the reference distributions for these tests. Section 4 illustrates the methods on two examples. Section 5 gives concluding remarks. Derivations are provided in an Appendix.

2. Notation

2.1. ZIP and Poisson-hurdle models

Consider a discrete random variable Y with a zero-inflated Poisson (ZIP) distribution:

$$\begin{aligned} \mathrm{P}(Y=0) &= \phi + (1-\phi)\exp(-\theta) \\ \mathrm{P}(Y=y) &= (1-\phi)\frac{\exp(-\theta)\theta^y}{y!} \quad \text{for } y = 1, 2, \ldots, \end{aligned}$$

where $0 < \phi < 1$ and $\theta > 0$. This standard parameterization represents the ZIP model as a mixture of a point mass at zero and a Poisson random variable with mean θ. Given that $\exp(-\theta) < 1$, the $\mathrm{P}(Y = 0)$ exceeds that obtained under a Poisson(θ) model. This leads to an excess number of zero counts and overdispersion relative to the Poisson model.

The ZIP model can be extended to $-\exp(-\theta)/\{1-\exp(-\theta)\} < \phi < 1$, but the interpretation of the model as a mixture does not apply when $\phi < 0$. Following Yau and Lee (2001), we set $p = \{1-\exp(-\theta)\}(1-\phi)$ and reparameterize the extended ZIP model as follows:

$$\begin{aligned} \mathrm{P}(Y=0) &= 1-p \\ \mathrm{P}(Y=y) &= \frac{p\,\theta^y\,\exp(-\theta)}{\{1-\exp(-\theta)\}\,y!} \quad \text{for } y = 1, 2, \ldots, \end{aligned}$$

or, equivalently, as

$$\mathrm{P}(Y=y) = \frac{\theta^y}{y!}(1-p)^{1-d}p^d c(\theta)^d \quad \text{for } y = 0, 1, \ldots,$$

where $c(\theta) = \exp(-\theta)/\{1-\exp(-\theta)\}$ and $d = 1_{(Y>0)}$ indicates whether $Y > 0$. In this version of the model, identified as ZIP(p, θ), the probability of a zero count is arbitrary ($0 < 1-p < 1$) and the conditional distribution of Y given $Y > 0$ follows a truncated Poisson(θ) distribution. For later reference, $\mu(\theta) = \theta/\{1-\exp(-\theta)\}$ and $\nu(\theta) = (\theta+\theta^2)/\{1-\exp(-\theta)\} - \mu(\theta)^2$ are the mean and variance of the truncated Poisson(θ) distribution. The ZIP(p, θ) model is commonly referred to as the Poisson-hurdle model in the econometrics literature (Cameron and Trivedi, 1998, p. 125).

The parameterization of the ZIP model is not critical when testing homogeneity of distributions. The ZIP(p, θ) version of the model is used in our analysis. However, the parameterization is important in regression settings where the parameters are modeled as functions of covariates; see Yau and Lee (2001). Section 5 discusses this issue. Lambert (1992), Gupta et al. (1996), Cameron and Trivedi (1998), and Böhning et al. (1999) give further details on ZIP models.

2.2. An exact test for homogeneity

Assume that we have independent random samples from k populations, with $Y_{i1}, \ldots, Y_{in_i} \sim \mathrm{ZIP}(p_i, \theta_i)$ for $i = 1, \ldots, k$. Let $\boldsymbol{Y_i} = (Y_{i1}, \ldots, Y_{in_i})'$ and $\boldsymbol{D_i} = (D_{i1}, \ldots, D_{in_i})'$, where $D_{ij} = 1_{(Y_{ij}>0)}$. Also, set $\boldsymbol{Y_+} = (Y_{1+}, \ldots, Y_{k+})'$ and $\boldsymbol{D_+} = (D_{1+}, \ldots, D_{k+})'$, where $Y_{i+} = \sum_j Y_{ij}$ and $D_{i+} = \sum_j D_{ij}$ are the sum of the counts and the number of positive counts in the ith group, respectively. Similarly, define $Y_{++} = \sum_i Y_{i+}$ and $D_{++} = \sum_i D_{i+}$. Finally, let $\hat{p}_i = D_{i+}/n_i$ be the proportion of positive counts in the ith group, and let $\hat{p} = D_{++}/n$ be the pooled proportion of positive counts, where $n = \sum_i n_i$.

The joint distribution of the vectors of counts for this k-group model is given by

$$\begin{aligned} \mathrm{P}(\boldsymbol{Y_1} = \boldsymbol{y_1}, \ldots, \boldsymbol{Y_k} = \boldsymbol{y_k}) &\equiv \mathrm{P}(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k}) \\ &= \prod_{i=1}^{k} (1-p_i)^{n_i - d_{i+}} p_i^{d_{i+}} c(\theta_i)^{d_{i+}} \frac{\theta_i^{y_{i+}}}{\prod_j y_{ij}!}. \end{aligned}$$

Note that

$$\mathrm{P}(\boldsymbol{D_1} = \boldsymbol{d_1}, \ldots, \boldsymbol{D_k} = \boldsymbol{d_k}) \equiv \mathrm{P}(\boldsymbol{d_1}, \ldots, \boldsymbol{d_k}) = \prod_{i=1}^{k} (1-p_i)^{n_i - d_{i+}} p_i^{d_{i+}}$$