JSPI: 5591 Model 3G pp. 1–21 (col. fig: NIL)

ARTICLE IN PRESS

Journal of Statistical Planning and Inference xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Analysis of multinomial counts with joint zero-inflation, with an application to health economics

Alpha Oumar Diallo [a,b], Aliou Diop [a], Jean-François Dupuy [b,*]

[a] LERSTAD, CEA-MITIC, Gaston Berger University, Saint Louis, Senegal
[b] IRMAR-INSA, Rennes, France

ARTICLE INFO

Article history:
Received 24 October 2016
Received in revised form 13 May 2017
Accepted 12 September 2017
Available online xxxx

Keywords:
Excess zeros
Health-care utilization
Multinomial logit

ABSTRACT

Zero-inflated regression models for count data are often used in health economics to analyze demand for medical care. Indeed, excess of zeros often affects health-care utilization data. Much of the recent econometric literature on the topic has focused on univariate health-care utilization measures, such as the number of doctor visits. However, health service utilization is usually measured by a number of different counts (*e.g.*, numbers of visits to different health-care providers). In this case, zero-inflation may jointly affect several of the utilization measures. In this paper, a zero-inflated regression model for multinomial counts with joint zero-inflation is proposed. Maximum likelihood estimators in this model are constructed and their properties are investigated, both theoretically and numerically. We apply the proposed model to an analysis of health-care utilization.

1. Introduction

Statistical modeling of count data with zero inflation has become an important issue in numerous fields and in particular, in econometrics. The zero inflation (or excess zeros) problem occurs when the proportion of zero counts in the observed sample is much larger than predicted by standard count models. In health economics, this issue often arises in analysis of health-care utilization, as measured by the number of doctor visits (Sarma and Simpson, 2006; Sari, 2009; Staub and Winkelmann, 2013). The present work is also motivated by an econometric analysis of health-care utilization and is illustrated by a data set described by Deb and Trivedi (1997).

Deb and Trivedi (1997) investigate the demand for medical care by elderlies in the United States. Their analysis is based on data from the National Medical Expenditure Survey (NMES) conducted in 1987 and 1988. These data provide a comprehensive picture of how Americans (aged 66 years and over) use and pay for health services. Six measures of health-care utilization were reported in this study, namely the number of visits to a doctor in an office setting, the number of visits to a non-doctor health professional (such as a nurse, optician, physiotherapist…) in an office setting, the number of visits to a doctor in an outpatient setting, the number of visits to a non-doctor in an outpatient setting, the number of visits to an emergency service and the number of hospital stays. A feature of these data is the high proportion of zero counts observed for some of the health-care utilization measures, that is, there is a high proportion of non-users of the corresponding health-care service over the study period. In addition to health services utilization, the data set also provides information on health status, sociodemographic characteristics and economic status. Deb and Trivedi (1997) analyze separately each measure of health-care utilization by fitting models for zero-inflated count data to each type of health-care usage in turns. However, several

* Corresponding author.
E-mail addresses: Alpha-Oumar.Diallo1@insa-rennes.fr (A.O. Diallo), aliou.diop@ugb.edu.sn (A. Diop), Jean-Francois.Dupuy@insa-rennes.fr (J.-F. Dupuy).

https://doi.org/10.1016/j.jspi.2017.09.005
0378-3758/

studies suggest that health-care utilization measures are not independent (Gurmu and Elder, 2000; Wang, 2003). Therefore, we suggest to analyze jointly the various health-care utilization measures by fitting a multinomial logistic regression model to the data.

For illustrative purpose, and in order to keep notations simple, we will illustrate our model and methodology by considering three out of the six measures of health-care utilization, namely the: (i) number Z_1 of consultations with a non-doctor in an office setting (denoted by *ofnd* in what follows), (ii) number Z_2 of consultations with a non-doctor in an outpatient setting (*opnd*) and (iii) number Z_3 of consultations with a doctor in an office setting (*ofd*). If m_i denotes the total number of consultations for the ith individual and $\mathbf{X}_i$ is a vector of covariates for this individual, we let $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ and we assume that Z_i has a multinomial distribution $\text{mult}(m_i, \mathbf{p}_i)$, where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$, $p_{1i} = \mathbb{P}(Z_{1i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofnd*, $p_{2i} = \mathbb{P}(Z_{2i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *opnd* and $p_{3i} = \mathbb{P}(Z_{3i} = 1|\mathbf{X}_i)$ is the probability that a consultation is of type *ofd*. We consider individuals in the NMES data set who have a total number of consultations less than or equal to 25. Among these 3224 individuals, frequencies of zero in variables *ofnd*, *opnd* and *ofd* are 62.7%, 81.3% and 1.5% respectively. Frequencies of zeros occurring simultaneously in variables of pairs (*ofnd* and *opnd*), (*ofnd* and *ofd*) and (*opnd* and *ofd*) are 51.7%, 0.24% and 1% respectively. That is, 51.7% of the surveyed subjects did not use any services associated with counts Z_1 and Z_2. This high frequency and the very low frequency of zero counts for *ofd* suggest that there may exist some permanent non-users of *ofnd* and *opnd*, *i.e.*, individuals who would never use these health-care services. In other words, there may exist an excess of observations of the form $(0, 0, m_i)$ in the data set.

To accommodate these observations, we propose to define, for each individual i, a zero-inflated multinomial regression model as the mixture

$$\pi_i \cdot \delta_{(0,0,m_i)} + (1 - \pi_i) \cdot \text{mult}(m_i, \mathbf{p}_i) \tag{1.1}$$

of the multinomial distribution $\text{mult}(m_i, \mathbf{p}_i)$ with a degenerate distribution $\delta_{(0,0,m_i)}$ at $(0, 0, m_i)$. π_i represents the probability that the ith individual is a permanent non-user of health-care services of the type *ofnd* and *opnd*.

Mixture models for zero-inflated count data date back to early '90s. Zero-inflated Poisson (ZIP) regression was proposed by Lambert (1992) and further developed by Dietz and Böhning (2000), Li (2011), Lim et al. (2006) and Monod (2014), among many others. Zero-inflated negative binomial (ZINB) regression was proposed by Ridout et al. (2001), see also Moghimbeigi et al. (2008), Mwalili et al. (2008) and Garay et al. (2011). Hall (2000) and Vieira et al. (2000) introduced the zero-inflated binomial (ZIB) regression model, see also Diop et al. (2016). But to the best of our knowledge, and although some related models can be found in Kelley and Anderson (2008) and Bagozzi (in press), the zero-inflated multinomial model (1.1) has not been yet considered. Kelley and Anderson (2008) (respectively Bagozzi, in press) propose a model for a discrete ordinal (respectively nominal) dependent variable with levels $\{0, 1, \ldots, J\}$ and zero-inflation. However, authors do not report any systematic investigation of their models (such as model identifiability or estimation). In the present paper, we aim at providing a rigorous study of model (1.1) that will serve as a basis for future application of the model to real-data problems. We derive maximum likelihood estimators of parameters π_i and $\mathbf{p}_i$, we establish their asymptotic properties (consistency and asymptotic normality) and we assess their finite-sample behavior using simulations. Then, we illustrate the model on the health-care utilization data set described above.

The remainder of the paper is organized as follows. In Section 2, we specify precisely the model and we address the estimation of π_i and $\mathbf{p}_i$. In Section 3, we report results of our simulation study. Section 4 describes the health-care data analysis. A conclusion and some perspectives are provided in Section 5. All proofs are postponed to an Appendix.

2. Zero-inflated multinomial regression model

In this section, we describe the zero-inflated multinomial (ZIM) regression model. We consider two cases: (i) π_i is fixed (that is, $\pi_i = \pi$ for every individual) and (ii) π_i depends on covariates. In Section 2.3, identifiability of the ZIM model and asymptotics of the maximum likelihood estimator are described for fixed π but results can be generalized to case (ii) without major difficulty. Moreover, for notational simplicity, we consider the case where the multinomial response Z_i has $K = 3$ outcomes. Proofs for a general K proceed similarly.

2.1. Model and estimation with fixed π

Let $(Z_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ be independent random vectors defined on the probability space $(\Omega, \mathcal{C}, \mathbb{P})$. For every i, we assume that given the total $Z_{1i} + Z_{2i} + Z_{3i} = m_i$, the multivariate response $Z_i = (Z_{1i}, Z_{2i}, Z_{3i})$ is generated from the model

$$Z_i \sim \begin{cases} (0, 0, m_i) & \text{with probability } \pi, \\ \text{mult}(m_i, \mathbf{p}_i) & \text{with probability } 1 - \pi, \end{cases} \tag{2.2}$$

where $\mathbf{p}_i = (p_{1i}, p_{2i}, p_{3i})$ and $p_{1i} + p_{2i} + p_{3i}$=1. This model reduces to the standard multinomial distribution (with three modalities, here) if $\pi = 0$, while $\pi > 0$ leads to simultaneous zero-inflation in the first two modalities. We model probabilities p_{1i}, p_{2i} and p_{3i} $(i = 1, \ldots, n)$ via multinomial logistic regression:

$$p_{1i} = \frac{e^{\beta_1^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \quad p_{2i} = \frac{e^{\beta_2^\top \mathbf{X}_i}}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}} \quad \text{and} \quad p_{3i} = \frac{1}{1 + e^{\beta_1^\top \mathbf{X}_i} + e^{\beta_2^\top \mathbf{X}_i}}, \tag{2.3}$$

Please cite this article in press as: Diallo A.O., et al., Analysis of multinomial counts with joint zero-inflation, with an application to health economics. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.09.005.