Computational Statistics and Data Analysis 53 (2009) 3717–3725

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A generalized Waring regression model for count data

J. Rodríguez-Avi *, A. Conde-Sánchez, A.J. Sáez-Castillo, M.J. Olmo-Jiménez, A.M. Martínez-Rodríguez

Department of Statistics and Operations Research, University of Jaén, Spain

ARTICLE INFO

Article history:
Received 14 May 2008
Received in revised form 19 March 2009
Accepted 20 March 2009
Available online 28 March 2009

ABSTRACT

A regression model for count data based on the generalized Waring distribution is developed. This model allows the observed variability to be split into three components: randomness, internal differences between individuals and the presence of other external factors that have not been included as covariates in the model. An application in the field of sports illustrates its capacity for modelling data sets with great accuracy. Moreover, this yields more information than a model based on the negative binomial distribution.

1. Introduction

The Poisson distribution is used to model count data where the occurrence of an event is random and the occurrence rate is the same for all individuals. However, this occurrence rate may differ across individuals: this is known as heterogeneity (Long, 1997). When this heterogeneity may be explained by quantifiable and observable characteristics of the individuals, the Poisson regression model (*PRM*) is valid, in which the mean is a function of the observed variables or covariates. So, the heterogeneity is modelled as a deterministic function of the explanatory variables (Winkelmann, 2003). This model is the most basic for count data and is characterized by the equality of conditional mean and variance (equidispersion).

However, it is well known that the variability of data often exceeds the Poisson variability (overdispersion). It may be explained in several ways (Xekalaki, 1983, 2004; Winkelmann, 2003), among others, by the existence of unobserved heterogeneity, that is, by the presence of unfixed occurrence rates at each level of the model covariates. If it is assumed that this occurrence rate follows a gamma distribution, the resulting model is the negative binomial regression model (*NBRM*) (Hinde and Demétrio, 1998; Poortema, 1999; Cameron and Trivedi, 1998; Long, 1997). This allows us to consider a new source of variation that differs from the observed covariates and from randomness and thereby an additional component to explain the variability.

Moreover, the unobserved heterogeneity may be due to internal differences across individuals and external factors that might also be included as covariates in the regression model if they could be observed. In the *NBRM* both sources of variation in the occurrence rate are jointly considered by means of a gamma distribution.

In accident theory the univariate generalized Waring distribution, *UGWD*, (Irwin, 1968; Xekalaki, 1983) is an extension of the negative binomial distribution that allows three sources of variation to be distinguished: randomness, which is inherent in any random phenomenon, and the two aforementioned sources of heterogeneity between individuals, the one due to external factors, that is, different accident risk exposures (liability), and the other due to internal factors pertaining to each individual, that is, personal differences that are not related to external factors (proneness). A more general distribution which also considers this partition of the variance is studied by Rodríguez-Avi et al. (2007).

* Corresponding address: Despacho B3-058 Campus Universitario de Jaén, 23071 Jaén, Spain. Tel.: +34 953212207; fax: +34 953212034.
E-mail address: jravi@ujaen.es (J. Rodríguez-Avi).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.03.013

This work describes a regression model with a *UGWD* as its underlying distribution. The main advantage of this model over the *NBRM* is that the former allows us to distinguish the part of the unobserved heterogeneity due to the internal factors inherent to each individual and that due to the external factors such as those covariates that influence the variability of data but that have not been included in the model because they cannot be observed or measurable. From now on and by analogy with the terminology used in accident theory, these parts of the variance will be called proneness and liability, respectively. The performance of both models has been compared by simulation methods.

An example in the field of sports is considered to illustrate the behaviour of the model. Specifically, the dependent variable is the number of goals scored by the footballers of the Spanish football league over several seasons and the covariates are the position of the footballers on the pitch and the final classification of the team. Moreover, the number of matches played by each footballer has been initially considered only as offset and subsequently also as regressor. The effect of the covariates is studied, the fit obtained is compared against a regression model based on a negative binomial distribution and the relative weight of the three sources of variation (randomness, liability and proneness) in the presence of the covariates is computed.

2. Negative binomial regression models

Let Y be the response variable of a count model. In a *PRM*, $Y|x \sim Poisson(\lambda_x)$, where λ_x is the mean of the response variable for the values of the covariates, $x' = (x_1, \dots, x_p)$. Obviously, there is equidispersion in each level of the covariates, that is, $Var(Y|x) = E(Y|x)$.

As has been stated, if $Y|x$ is overdispersed, a way of explaining this excess variability is to propose a parametric model for λ_x. When $\lambda_x \sim Gamma(a_x, v_x)$, that is to say,

$$f(\lambda_x) = \frac{1}{v_x^{a_x}\Gamma(a_x)}\lambda_x^{a_x-1}e^{-\lambda_x/v_x}, \quad \lambda_x > 0, a_x, v_x > 0.$$

$Y|x$ has a negative binomial distribution with probability mass function (p.m.f.)

$$f(y|x) = \frac{\Gamma(a_x+y)}{\Gamma(a_x)y!}\left(\frac{1}{1+v_x}\right)^{a_x}\left(\frac{v_x}{1+v_x}\right)^{y}, \quad y = 0, 1, 2, \dots, a_x, v_x > 0, \tag{1}$$

denoted by $Y|x \sim NB(a_x, p_x)$ with $p_x = (1+v_x)^{-1}$. In this case, an *NBRM* arises. It should be emphasized that the model about λ_x is related to the unexplained heterogeneity, independently of its origin. In this model it verifies that

$$E(Y|x) = E(E(Y|x, \lambda_x)) = E(\lambda_x) = \mu_x = a_x v_x.$$

Different *NBRM* can be generated by linking μ_x, a_x and v_x with the explanatory variables. One of the most usual in data processing is given by

$$a_x = a, \qquad \mu_x = e^{\beta_0 + x'\beta},$$

with $\beta' = (\beta_1, \dots, \beta_p)$, where a does not depend on the covariates but does v_x. This model is known as *Negbin* II (Cameron and Trivedi, 1986) and it establishes a linear variance-mean rate:

$$\begin{aligned} Var(Y|x) &= E(Var(Y|x, \lambda_x)) + Var(E(Y|x, \lambda_x)) \\ &= E(\lambda_x) + Var(\lambda_x) = \mu_x + \frac{1}{a}\mu_x^2 = \mu_x\left(1 + \frac{1}{a}\mu_x\right). \end{aligned}$$

The first term represents the variability due to randomness and the second to differences between individuals. The partition of the variance for this model appears in Table 1. It can be observed that the variance rate due to heterogeneity across individuals tends to 1 as μ_x increases, whereas the variance rate due to randomness tends to 0. This means that, as the average number of occurrences of an event increases, the observed variability is more due to the individual heterogeneity than to randomness.

It should be pointed out that if $a \to \infty$ and $v_x \to 0$ with μ_x constant, the *Negbin* II model tends to the *PRM*, so the latter is nested within the former.

On the other hand, if v does not depend on the covariates but does a_x, that is, $v_x = v$ and $\mu_x = e^{\beta_0 + x'\beta}$, the *Negbin* I model appears (Cameron and Trivedi, 1986) in which the variance-mean rate is constant:

$$Var(Y|x) = (1+v)\mu_x.$$

In order to make comparisons, we focus on the *Negbin* II model, since it provides better fits for data included here than does the *Negbin* I model.

The NBRM recently appears within more general frameworks to model count data variables, as in Rigby et al. (2008), where all the distribution parameters can be modelled as functions of explanatory variables, or in Cordeiro et al. (2009), where a new class of discrete generalized nonlinear models is introduced.