Journal of Statistical Planning and Inference 145 (2014) 190–203

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

On generalized multinomial models and joint percentile estimation

I. Das, S. Mukhopadhyay *

Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India

ARTICLE INFO

Article history:
Received 29 November 2012
Received in revised form
21 August 2013
Accepted 23 August 2013
Available online 12 September 2013

Keywords:
Confidence regions
Multicategorical logistic link
Parameter orthogonality
Standardization

ABSTRACT

This article proposes a family of link functions for the multinomial response model. The link family includes the multicategorical logistic link as one of its members. Conditions for the local orthogonality of the link and the regression parameters are given. It is shown that local orthogonality of the parameters in a neighbourhood makes the link family location and scale invariant. Confidence regions for jointly estimating the percentiles based on the parametric family of link functions are also determined. A numerical example based on a combination drug study is used to illustrate the proposed parametric link family and the confidence regions for joint percentile estimation.

1. Introduction

In this article we address two issues related to multinomial response models, (i) a family of link functions and (ii) percentile estimation under a parametric link family. In the first few sections we propose a family of link functions for multinomial nominal response models. When working with multinomial data sets the common practice is to fit the multicategory logistic link function (Agresti, 2002, pp. 267–274). However, Czado and Santner (1992) show that if the link function is incorrectly assumed then it leads to biased estimates thus increasing the mean squared error of prediction. Using a data set based on a combination drug therapy experiment we show that parameter estimation is improved by using the proposed link family instead of the commonly used multivariate logistic link. The parametric link family proposed includes the multivariate logistic link as one of its members. In the later part of the article we discuss three methods for finding confidence regions for the percentiles of a multinomial response model. The confidence regions determined are based on the estimated values of the link parameters.

In univariate generalized linear models (GLMs), especially for binary data, family of link functions have been discussed by many researchers. Some of the one and two parameter link families for binary models are proposed by Prentice (1975, 1976); Pregibon (1980); Guerrero and Johnson (1982); Aranda-Ordaz (1981); Stukel (1988); Czado and Santner (1992); Czado (1992, 1993, 1997); Lang (1999). However, unlike binary regression models research papers on link families for multinomial responses are rarely found in the literature. The two parametric link families proposed by Genter and Farewell (1985) and Lang (1999) are only applicable to multinomial data sets with ordered categories. Till date we were unable to find any work which addresses a family of link functions for multinomial data sets with nominal responses. The situation is similar for percentile estimation methods in multinomial response models. Though a huge number of research papers (Hamilton, 1979; Carter et al., 1986; Williams, 1986; Huang, 2001; Biedermann et al., 2006; Li and Wiens, 2011) have been written on percentile estimation and effect of link misspecification on percentile estimation for binary data, almost no work has been done in the case of multinomial data. There are, however, many experimental situations where multinomial responses may be observed for each setting of a group of control

* Corresponding author. Tel.: +91 2225767495.
E-mail addresses: ishapathik@math.iitb.ac.in (I. Das), siuli@math.iitb.ac.in (S. Mukhopadhyay).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.08.015

variables. As a typical example we may consider a drug testing experiment, where both the efficacious and toxic responses of the drug/s are measured on the subjects. This results in two responses, efficacy and toxicity of the drug, both of which are binary in nature. Since the responses come from the same subject they are assumed to be correlated, and can be modeled using a multinomial distribution (Mukhopadhyay and Khuri, 2008). In this situation it may be of interest to the experimenter to jointly estimate the 100p percentile of the efficacy and toxic responses. In this article we discuss a numerical example based on the pain relieving and toxic effects of two analgesic drugs and determine confidence regions for the 100p percentiles of both the responses.

While parametric link families are able to improve the maximum likelihood fit when compared to canonical links, any correlation between the link and the regression parameters leads to an increase in the variances of the parameter estimates (Czado, 1997). However, it can be shown that if the parameters are orthogonal to each other then the variance inflation reduces to zero for large sample sizes (Cox and Reid, 1987). Conditions for local orthogonality in a neighbourhood were proposed by Czado (1997) for univariate GLMs. In this article we extend these conditions so that we can apply them to a multiresponse situation. It is also shown that the local orthogonality of the parameters imply location and scale invariance of the family of link functions.

The remainder of the article is organized as follows: In Section 2 we describe the family of link functions for the multinomial model. Detailed conditions of local orthogonality between the link and the regression parameters are given in Section 3. In Section 4 we discuss three interval methods for percentile estimation in a multinomial model. Using simulations we compare the three interval estimation methods using coverage probabilities. The proposed link family and confidence regions are illustrated with a numerical example based on a drug testing experiment in Section 5. We also propose a bootstrap interval in this section. Concluding remarks are given in Section 6.

2. A family of link functions for multinomial data

In this section the multinomial response model with a parametric link function is introduced. We use a scaled version of the multinomial distribution. The following three components are used to describe it:

- Distributional component: A random sample of size n, $\mathbf{y}_1,\dots,\mathbf{y}_n$, is selected from a multinomial distribution with parameters $(\boldsymbol{\pi}_i, n_i)$; $\boldsymbol{\pi}_i=(\pi_{i1},\dots,\pi_{iq})$, $i=1,\dots,n$. The density function of $\overline{\mathbf{y}}_i=\mathbf{y}_i/n_i$ also called the scaled multinomial distribution (Fahrmeir and Tutz, 2001, p. 76) is

$$s(\overline{\mathbf{y}}_i|\boldsymbol{\theta}_i,\phi,\omega)=\exp\left\{\frac{[\overline{\mathbf{y}}_i'\boldsymbol{\theta}_i-b(\boldsymbol{\theta}_i)]}{\phi}\omega_i+c(\mathbf{y}_i,\phi,\omega_i)\right\}, \tag{2.1}$$

  where

$$\boldsymbol{\theta}_i=\left[\log\left(\frac{\pi_{i1}}{1-\sum_{j=1}^{q}\pi_{ij}}\right),\dots,\log\left(\frac{\pi_{iq}}{1-\sum_{j=1}^{q}\pi_{ij}}\right)\right]', \quad b(\boldsymbol{\theta}_i)=-\log\left(1-\sum_{j=1}^{q}\pi_{ij}\right),$$
$$c(\mathbf{y}_i,\phi,\omega_i)=\log\left(\frac{n_i!}{y_{i1}!\dots y_{iq}!(n_i-y_{i1}-\dots-y_{iq})!}\right),$$

  $\omega_i=n_i$ and $\phi=1$. The total number of observations is $N=\sum_{i=1}^{n}n_i$.
- Linear predictor: A q dimensional linear predictor, $\boldsymbol{\eta}(\mathbf{x})=\mathbf{Z}(\mathbf{x})\boldsymbol{\beta}$, where $\mathbf{Z}(\mathbf{x})=\oplus_{j=1}^{q}\mathbf{f}_j(\mathbf{x})$, $\mathbf{f}_j(\mathbf{x})$ is a known vector function of $\mathbf{x}$, $\boldsymbol{\beta}=[\boldsymbol{\beta}_1',\dots,\boldsymbol{\beta}_q']'$ is the $p\times 1$ vector of unknown parameters with the jth component, $\boldsymbol{\beta}_j$, of length p_j and $p=\sum_{j=1}^{q}p_j$.
- Parametric link function: $\boldsymbol{\mu}=\boldsymbol{\pi}=\mathbf{h}(\boldsymbol{\alpha},\boldsymbol{\eta})$, where $\mathbf{h}(\boldsymbol{\alpha},\cdot)=[h_1(\boldsymbol{\alpha},\cdot),\dots,h_q(\boldsymbol{\alpha},\cdot)]'$, $\boldsymbol{\alpha}_{r\times 1}=[\boldsymbol{\alpha}_1',\dots,\boldsymbol{\alpha}_q']'$, $\boldsymbol{\alpha}_j$ is of length r_j and $\sum_{j=1}^{q}r_j=r$.

2.1. Proposed form of parametric link function

Several researchers (Stukel, 1988; Czado, 1989) propose the following generalization for binary response models with a logistic link function:

$$\mu(\mathbf{x})=E(y|\mathbf{x})=h(\boldsymbol{\alpha},\eta)=\frac{\exp\{G(\boldsymbol{\alpha},\eta)\}}{[1+\exp\{G(\boldsymbol{\alpha},\eta)\}]},$$

where $G(\boldsymbol{\alpha},\cdot)$ is a generating family with the unknown link parameter $\boldsymbol{\alpha}$. For example using the generating family by Czado (1989) we get

$$G(\boldsymbol{\alpha},\eta)=\begin{cases}\dfrac{(1+\eta)^{\alpha_1}-1}{\alpha_1} & \text{if } \eta\geq 0\\[2ex] -\dfrac{(1-\eta)^{\alpha_2}-1}{\alpha_2} & \text{if } \eta<0,\end{cases}$$