HOW TO DEFINE DEVIANCE RESIDUALS IN MULTINOMIAL REGRESSION

Giovanni Romeo [1], Mariangela Sciandra[1] and Marcello Chiodi[1]

[1] Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università degli Studi di Palermo, (e-mail: giovanni.romeo1989@gmail.com, mariangela.sciandra@unipa.it)

ABSTRACT: This work is devoted to the study of diagnostic tools for categorical data models with an emphasis on the presence of continuous covariates. In particular, the aim is to define a new class of residuals from the parametric multinomial family of models and to study their asymptotics properties. In logistic regression (as in generalized linear models), there are a few different kinds of residuals; we propose a generalization of deviance residuals as defined in logistic regression to the multinomial case and propose their use in order to identify inadequacies in a multinomial model.

KEYWORDS: multinomial response models, diagnostic tools, extended deviance residuals.

1 Introduction

Multinomial logistic regression is often defined as a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to derive the probability categorical membership and requires diagnostic methods based on residuals analysis in order to evaluate the model adequacy.
As it is known, the residual vector from the linear model $y \sim N(X\beta, \sigma^2 I_n)$ with X an $n \times p$ matrix of rank p, is given by $r = y - \hat{y}$, where $\hat{y} = X\hat{\beta}$ and $\hat{\beta}$ is the last squares estimator of β. If the model is correct, the n residuals n_i have zero means, variance matrix $\sigma^2(I-P)$, where P is the projection matrix $X(X'X)^{-1}X'$, and they are uncorrelated with the fitted values $\hat{y}_i$. When non linear models are fitted, residuals do not yet share these useful properties. So, in order to overcome these problems one possibility is to use the so-called *adjusted residuals* (Haberman,1973) or the *projected residuals* as defined in Cook & Tsai (1985). Unfortunately, these diagnostics are valid only under the requirement of reasonably large sample sizes. One problem of using such

methods is the lack of guidelines about the right sample size necessary to warrant their use. In absence of such guidelines, the validity of these methods may be questionable and alternative residuals which do not depend on this requirement may be used. Aim of this paper is to define a new class of residuals for multinomial regression which behave as much as possible like linear regression residuals. So, after a brief remind on multinomial regression models, in the third section a new definition of deviance residuals in multinomial regression is proposed; in the fourth section an application to neurological data is presented to deal with a new multiple diagnostic plot. Some ideas for future work conclude this paper.

2 The multinomial logistic regression model

Multinomial logistic regression is a generalization of classical logistic regression to the polytomous case. Let J be the number of discrete categories of the dependent variable and Z be a random variable that can take on one of J possible values. By aggregating the data into populations each of which represents one unique combination of independent variable settings, a column vector $\mathbf{m}$ will contain elements m_i which represent the number of observations in population i, and such that $\sum_{i=1}^{M} m_i = n$, the total sample size. Then, if observations are independent, then each Z_i is a multinomial random variable. Let $\mathbf{y}$ be a matrix with M rows and $J-1$ columns, y_{ij} represents the observed count of the $j-$th value of Z_i. Similarly, π is a matrix where π_{ij} represents the probability of observing the $j-$th value in the $i-$th population.

Let $\mathbf{X}$ be the $M-$rows design matrix of K explanatory variables. Let β be a matrix with $K+1$ rows and $J-1$ columns, such that each element β_{kj} contains the parameter estimate for the $k-$th covariate and the $j-$th value of the dependent variable. Assuming J to be the baseline, the multinomial logistic regression model can be derived by equating the linear component to the log of the odds of a generic $j-$th observation compared to the $J-$th observation:

$$log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj}$$

with $i = 1,2,\ldots,M$ and $j = 1,2,\ldots,J-1$. So, the model will consist of $J-1$ logit equations that are fitted simultaneously. As a consequence, $J-1$ residuals can be drawn for each multinomial observation. This increase in dimensionality complicates residual analyses because a generalization of classical residual tools is required. The most simple solution consists in creating a sequence

of binary residuals by collapsing response categories. Nevertheless, resulting binary residuals, evaluated for each threshold used in dichotomization, will be correlated and a possible confounding could derive (Seber and Nyangoma, 2000).

3 Deviance residuals in multinomial models: a proposal

In order to overcome problems related to residuals for polytomous response data, we propose an extension of standard deviance residuals (McCullagh and Nelder, 1999). Starting from the multinomial log-likelihood:

$$Log(\pi; y) = \sum_{i=1}^{M} \sum_{j=1}^{J} y_{ij} log(\pi_{ij})$$

The deviance can be derived as $G = 2\sum_{i=1}^{M} \sum_{j=1}^{J} y_{ij} log \frac{y_{ij}}{\hat{\pi}_{ij}}$.
Then, multinomial *deviance contributions* d_i can be defined as:

$$d_i = -2 \sum_{j=1}^{J} y_{ij} log \hat{\pi}_{ij} \quad i = 1, \ldots, M \tag{1}$$

with $\sum_{i=1}^{M} d_i = G$. Given the polytomous nature of the response variable, the extension of the definition of classical deviance residuals asks for a criterion to univocally assign signs to these individual contributions. We propose to define *signed extended deviance residuals* as below:

$$_d r^*_{ij} = s_{ij} \sqrt{d_i} \quad \forall i = 1, \ldots, M \quad and \quad j = 1, \ldots, J-1 \tag{2}$$

where $s_{ij} = 1$ if $y_{ij} = 1$, $s_{ij} = -1$ if $y_{iJ} = 1$. It is important to outline that for each i belonging to the j-th category, r_{ij} will be not defined out of categories j and J. In terms of graphical representation we propose to represent $J-1$ scatterplots having on the x-axis the linear predictor $\hat{\eta}_{ij}$ for the j-th response category and on the y-axis the corresponding j-th deviance residual. This representation, as standard logit do, allows a graphical comparison of residuals for the generic category j with those from the baseline J on the $j-$th linear predictor scale and potentially should emphasize lack of fit.

4 A real data example

This data are from the CogItA (Cognitive Impairment through Aging) project, a large hospital-based prospective study started in January 1999 and still ongoing. We selected a large sample (2915) of patients affected by Mild Cognitive

Impairment (MCI), followed by the Department of Neurology and Cognitive Disorders of the University of Palermo. Aim of the study was to analyze differences in four different types of MCI depending on a set of covariates (age, sex, education, vascular risk factors and brain lesions). A multinomial model was fitted with the "neurologically health" subjects as reference category.

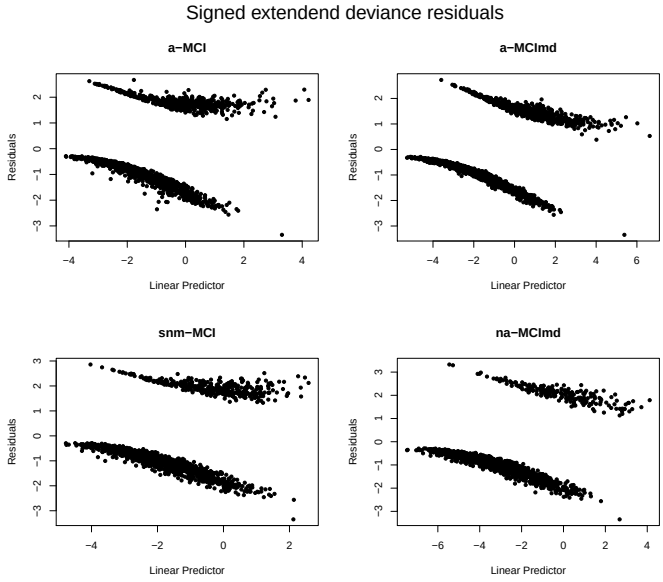

Figure 1. *Signed extended deviance residuals from a fitted multinomial model for the Mild Cognitive Impairment ("neurologically health" subjects are the reference category).*

In Fig.1 *signed extended deviance residuals* from the fitted models computed with (2) are plotted. In particular, residuals for the first and the third group (a-MCI and snm-MCI) show an increasing variance with the linear predictors, suggesting some overdispersion in the data.

The plot allows to derive specific comparisons between the reference level and the other response categories and as a consequence it makes possible to detect where the model needs for improvements. Properties of this new class of residuals will be studied also using simulations.

References

COOK, RD, & TSAI, CHIH-LING. 1985. Residuals in nonlinear regression. *Biometrika*, **72**(1), 23–29.

HABERMAN, SHELBY J. 1973. The analysis of residuals in cross-classified tables. *Biometrics*, 205–220.

MCCULLAGH, PETER, NELDER, JOHN A, & MCCULLAGH, P. 1989. *Generalized linear models*. Vol. 2. Chapman and Hall London.

SEBER, GAF, & NYANGOMA, SO. 2000. Residuals for multinomial models. *Biometrika*, **87**(1), 183–191.