Computational Statistics and Data Analysis 103 (2016) 17–27

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On a dispersion model with Pearson residual responses

K.Y.K. Wu *, W.K. Li

Department of Statistics & Actuarial Science, University of Hong Kong, Pok Fu Lam, Hong Kong

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form 25 April 2016
Accepted 30 April 2016
Available online 6 May 2016

Keywords:
Pearson residual
Pseudo-likelihood function
Dispersion model

ABSTRACT

Dispersion regression is often used to predict the expected deviance in a generalised linear model. Using the individual deviance residual as the response variable in that model is considered the standard approach in dispersion modelling. In this paper, we investigate an alternative approach by fitting the dispersion model on the individual Pearson residual responses, which is more straightforward than and has superior interpretability to the deviance approach because no transformation on the observed and expected responses via the likelihood function is required. However, the mean and dispersion parameters are non-orthogonal if the model parameter estimates are obtained by maximising the pseudo-likelihood function. Consequently, the mean and dispersion regression parameters must be estimated simultaneously, and the estimation algorithm is multidimensional and hence more complex. As the asymptotic behaviour of both the deviance and Pearson residuals suggests that they should converge, we expect Pearson residual dispersion models to perform in the same way as or even better than deviance residual models.

1. Introduction

Since it was first proposed by Wedderburn (1974), the quasi-likelihood function, based on which maximum likelihood (ML) estimates of the dispersion model can be determined, has generally been considered the orthodox approach in dispersion modelling. The deviance residual resulting from the quasi-likelihood function has therefore generally been considered the response variable in the dispersion model. However, as a goodness-of-fit measure of the model, the deviance residual is sometimes lacking in interpretability because of its dependence on the likelihood function.

Other residuals such as the Pearson or Anscombe residual have been suggested in the literature as alternatives to the deviance residual. However, only a few researchers have devoted attention to them, and their potential in dispersion modelling has never been studied in detail or completely exploited, although McCullagh and Nelder (1989) briefly discussed it. Owing to its definition, the Pearson residual has a higher degree of interpretability as it measures the standardised distance between an observed and expected response directly, and is therefore mostly used primarily for the goodness of fit of contingency tables as the cell frequency is usually considered as the Poisson count.

In this paper, we show how individual Pearson residuals can be used as response variables in a double generalised linear model (GLM) of mean and dispersion. As demonstrated in later sections, the estimation algorithm becomes more complex because the mean and dispersion parameters are non-orthogonal when maximising a pseudo-likelihood (PL) function (Carroll and Ruppert, 1988) in which individual Pearson residuals represent the response variables. As proposed by Smyth (1989) and Smyth and Verbyla (1996), simultaneous estimation of the double mean and dispersion model is needed if the parameters are non-orthogonal. Joint estimation of the regression parameters of both submodels, however, leads to

* Corresponding author.
E-mail addresses: kykarlwu@gmail.com (K.Y.K. Wu), hrntlwk@hku.hk (W.K. Li).

http://dx.doi.org/10.1016/j.csda.2016.04.015
0167-9473/

two-dimensional score functions and a highly complex information matrix structure, which can be simplified only in special circumstances. By contrast, the asymptotic properties of the ML estimates obtained from the PL function are the same as those derived from the quasi-likelihood function, and we thus expect our approach to provide a real alternative method to dispersion modelling.

By means of simulation studies, we investigate the features of the ML estimates and the goodness of fit of the models on Poisson- and Gaussian-distributed responses. The exploitation of individual Pearson residuals as a useful analytic tool and the efficiency of that tool in detecting underestimated standard errors form a significant part of our research. In particular, we compare the results with models in which the overdispersion feature is ignored. The applicability of our model is examined via a case study example given in a later section of the paper.

2. Generalised linear dispersion model

Let $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ be an n-dimensional random vector whose underlying distribution belongs to the exponential family. The density of the ith observation is given by

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\}, \tag{1}$$

where θ_i is the canonical, and ϕ the dispersion parameter. In our model formulation, we consider the special case of the function $a(\phi) = \phi/w_i$, where w_i is the weight of the ith observation, which is usually equal to 1. The cumulant function denoted by $b(\cdot)$ satisfies the relationship $\partial b/\partial\theta_i = \dot{b}(\theta_i) = \mu_i$ and $\partial^2 b/\partial\theta_i^2 = \ddot{b}(\theta_i) = V(\mu_i)$, the variance function of y_i, whereas the form of $c(\cdot)$ depends on the distribution of y_i.

Let the row vector $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{ik}\}$ be the ith row of an $n \times k$ matrix $\boldsymbol{X}$. $\mathbb{E}(y_i \mid \boldsymbol{x}_i) = \mu_i$ is the conditional expectation of y_i predicted by a GLM (Nelder and Wedderburn, 1972):

$$\mathbb{E}(y_i \mid \boldsymbol{x}_i) = \mu_i = g^{-1}(\boldsymbol{x}_i\boldsymbol{\beta}), \tag{2}$$

where $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_k\}^\intercal$ is the unknown regression coefficient vector to be estimated and $g^{-1}(\cdot)$ is the inverse function of an at least twice differentiable link function g to map the linear predictor $\eta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ to the scale of μ_i. For instance, if Y is Poisson-distributed, the natural choice for g^{-1} would be exponential, as μ_i is positive. For a Gaussian-distributed Y, we have the special case of a general linear regression where g is identical. Model (2) is called the *mean submodel* in the following.

The generalised Pearson statistic r_P^2 is widely used as a goodness-of-fit measure, particularly for contingency tables. It is defined as the standardised complete sample deviance of the expected to the observed response:

$$r_p^2 = \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)^2}{V(\mu_i)},$$

where $V(\mu_i)$ is the aforementioned variance function with

$$\mathrm{Var}(y_i) = \frac{\phi}{w_i}V(\mu_i). \tag{3}$$

Note that each term in the summation measures the individual squared Pearson residual $r_{P_i}^2$ with asymptotic response mean ϕ. As the definition suggests, the Pearson residual is a straightforward measure of the goodness of fit of $\hat{\mu}_i$. In contrast, goodness of fit based on the individual deviance residual, which is defined as

$$D_i(y_i, \mu_i) = 2[\ell(y_i, y_i) - \ell(y_i, \mu_i)] = 2w_i\int_{\mu_i}^{y_i}\frac{y_i - t_i}{V(t_i)}dt_i, \tag{4}$$

depends either on evaluation of the likelihood function

$$\ell(y_i, \mu_i) = \sum_{i=1}^{n}\left\{\frac{w_i}{\phi}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\}$$

at y_i and μ_i or on the integration of the quasi-likelihood function, as the last equality sign in Eq. (4) suggests. Depending on the choice of $g(\cdot)$, and consequently $g^{-1}(\cdot)$, the range of μ_i may not conform to the domain of $b^{-1}(\cdot)$. The Pearson residual is therefore advantageous in terms of its interpretability and computational convenience.

In general, we suppose for all i that the dispersion parameter ϕ is constant in the sense that we use merely one value to describe the dispersion of the entire sample. If Y_i is continuous, the underlying distributions such as normal or gamma distributions contain two parameters, whereas discrete distributions usually have only one. As a result, ϕ can be estimated from the sample for Gaussian or gamma responses, whereas count and binary responses are assumed to have a dispersion equal to 1.

However, homoscedasticity can actually be rarely found in empirical studies, as well as satisfying the relationship amongst μ_i, ϕ and $\mathrm{Var}(y_i)$, as defined in (3), in reality, and the theoretical dispersion often exceeds the empirical one.