Computational Statistics and Data Analysis 68 (2013) 44–51

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On diagnostics in double generalized linear models

Gilberto A. Paula *

Instituto de Matemática e Estatística - USP, Brazil

ARTICLE INFO

Article history:
Received 19 December 2012
Received in revised form 6 June 2013
Accepted 6 June 2013
Available online 14 June 2013

Keywords:
Deviance component residual
Double gamma model
Leverage measure
Local influence
Pearson residual
Residual analysis

ABSTRACT

The aim of this paper is to propose some diagnostic methods in double generalized linear models (DGLMs) for large samples. A review of DGLMs is given, including the iterative process for the estimation of the mean and precision coefficients as well as some asymptotic results. Then, a variety of diagnostic tools, such as leverage measures and curvatures of local influence under some usual perturbation schemes, the standardized deviance component, and Pearson residuals, are proposed. The diagnostic plots are constructed for the mean and precision models, and an illustrative example, in which the texture of four different forms of light snacks is compared across time with the texture of a traditional one, is analyzed under appropriate double gamma models. Some of the diagnostic procedures proposed in the paper are applied to analyze the fitted selected model.

1. Introduction

The class of double generalized linear models (DGLMs) was proposed by Smyth (1989), and Verbyla (1993) derived some case deletion diagnostics for linear heteroscedastic models under maximum likelihood (ML) and restricted maximum likelihood (REML) estimation. The REML method has been considered more reliable than ML for small samples (Smyth and Verbyla, 1999), and various papers have been published under this methodology. For example, Smyth and Verbyla (1999) investigated the sensitivity of the restricted maximum likelihood estimates (REMLEs) for some DGLMs, whereas Smyth and Jørgensen (2002) applied the framework of DGLMs to insurance claims. However, under the ML approach, little has been done on diagnostic methods. In this paper, some usual diagnostic quantities, such leverage measures, local influence curvatures, and Pearson and deviance component residuals are derived for DGLMs under ML. A large sample data set, in which the texture of five snack types is compared across time, is fitted under appropriate double gamma models, and a diagnostic analysis is performed with the quantities proposed in the paper to analyze the selected fitted model.

The paper is organized as follows. In Section 2, a review of DGLMs is presented, whereas in Section 3 we derive some useful diagnostic quantities, such as generalized leverages, curvatures of local influence under some usual perturbation schemes, and standardized forms for the Pearson and deviance component residuals. All the calculations are performed for the mean and precision models. The application is given in Section 4, and Section 5 deals with some conclusions. Approximate standardized forms for the Pearson residuals are derived in the Appendix.

2. Review of DGLMs

Let $Y_1, \ldots, Y_n$ be independent random variables with the density function of Y_i expressed in the exponential family form,

$$f(y_i; \theta_i, \phi_i) = \exp[\phi_i\{y_i\theta_i - b(\theta_i)\} + c(y_i; \phi_i)], \tag{1}$$

where $c(y_i, \phi_i) = d(\phi_i) + \phi_i a(y_i) + u(y_i)$ (normal, inverse Gaussian, and gamma distributions), $b(\cdot)$, $d(\cdot)$, $a(\cdot)$, and $u(\cdot)$ are

* Correspondence to: Departamento de Estatística, IME-USP, Rua do Matão 1010, Cidade Universitária, 05508-090, São Paulo-SP, Brazil. Tel.: +55 1130916239.

E-mail addresses: giapaula@ime.usp.br, gilbertop056@gmail.com.

0167-9473/$ – see front matter
http://dx.doi.org/10.1016/j.csda.2013.06.008

Table 1
Useful quantities derived for some exponential family distributions.

	Normal	Inverse Gaussian	Gamma
t_i	$y_i\mu_i - \frac{1}{2}(\mu_i^2 + y_i^2)$	$-\{y_i/2\mu_i^2 + \mu_i^{-1} + (2y_i)^{-1}\}$	$\log(y_i/\mu_i) - y_i/\mu_i$
$d(\phi)$	$\frac{1}{2}\log\phi$	$\frac{1}{2}\log\phi$	$\phi\log\phi - \log\Gamma(\phi)$
$d'(\phi)$	$(2\phi)^{-1}$	$(2\phi)^{-1}$	$(1+\log\phi) - \psi(\phi)$
$d''(\phi)$	$-(2\phi^2)^{-1}$	$-(2\phi^2)^{-1}$	$\phi^{-1} - \psi'(\phi)$

$\Gamma(\cdot)$, $\psi(\cdot)$, and $\psi'(\cdot)$ denote the gamma, digamma, and trigamma functions.

twice differentiable functions, θ_i is the canonical parameter, and ϕ_i (ϕ_i^{-1}) is the precision (dispersion) parameter. Alternatively, taking $T_i = Y_i\theta_i - b(\theta_i) + a(Y_i)$, one may express the density function of T_i (given θ_i) in the exponential family form (1), namely

$$f(t_i;\phi_i) = \exp\{\phi_i t_i + d(\phi_i) + u(y_i)\}. \tag{2}$$

From standard regularity conditions it follows that $\mu_i = \mathrm{E}(Y_i) = b'(\theta_i)$ and $\mathrm{Var}(Y_i) = \phi_i^{-1}V(\mu_i)$, where $V(\mu_i) = V_i = b''(\theta_i)$ is the variance function, $\mathrm{E}(T_i) = -d'(\phi_i)$ and $\mathrm{Var}(T_i) = -d''(\phi_i)$. Table 1 presents some of the quantities above derived for the normal, inverse Gaussian, and gamma distributions.

The DGLMs are defined by assuming the systematic components $g(\mu_i) = \eta_i = \mathbf{x}_i^\top\boldsymbol{\beta}$ and $h(\phi_i) = \lambda_i = \mathbf{z}_i^\top\boldsymbol{\gamma}$, where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^\top$ and $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_q)^\top$ are the model parameters to be estimated, $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^\top$ and $\mathbf{z}_i = (z_{i1},\ldots,z_{1q})^\top$ contain values of explanatory variables, and $g(\cdot)$ and $h(\cdot)$ are the link functions. Models (1) and (2), called the mean model and the precision model, respectively, belong to the class of generalized additive models for location, scale, and shape proposed by Rigby and Stasinopoulos (2005).

2.1. Parameter estimation

The score function for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ may be, respectively, expressed as

$$\mathbf{U}_\beta = \mathbf{X}^\top\boldsymbol{\Phi}\mathbf{W}^{1/2}\mathbf{V}^{-1/2}(\mathbf{y}-\boldsymbol{\mu}) \quad \text{and} \quad \mathbf{U}_\gamma = \mathbf{Z}^\top\mathbf{H}_\gamma^{-1}(\mathbf{t}-\boldsymbol{\mu}_T),$$

where $\mathbf{X}$ is an $n\times p$ matrix of rows $\mathbf{x}_i^\top$ $(i=1,\ldots,n)$, $\mathbf{W} = \mathrm{diag}\{\omega_1,\ldots,\omega_n\}$ with weights $\omega_i = (d\mu_i/d\eta_i)^2/V_i$, $\mathbf{V} = \mathrm{diag}\{V_1,\ldots,V_n\}$, $\boldsymbol{\Phi} = \mathrm{diag}\{\phi_1,\ldots,\phi_n\}$, $\mathbf{y} = (y_1,\ldots,y_n)^\top$, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^\top$, $\mathbf{Z}$ is an $n\times q$ matrix of rows $\mathbf{z}_i^\top$ $(i = 1,\ldots,n)$, $\mathbf{H}_\gamma = \mathrm{diag}\{h'(\phi_1),\ldots,h'(\phi_n)\}$, $\mathbf{t} = (t_1,\ldots,t_n)^\top$, and $\boldsymbol{\mu}_T = (\mathrm{E}(T_1),\ldots,\mathrm{E}(T_n))^\top = (-d'(\phi_1),\ldots,-d'(\phi_n))^\top$. The Fisher information matrices for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are, respectively, given by

$$\mathbf{K}_{\beta\beta} = \mathbf{X}^\top\boldsymbol{\Phi}\mathbf{W}\mathbf{X} \quad \text{and} \quad \mathbf{K}_{\gamma\gamma} = \mathbf{Z}^\top\mathbf{P}\mathbf{Z},$$

where $\mathbf{P} = \mathrm{diag}\{p_1,\ldots,p_n\}$ with $p_i = -d''(\phi_i)\{h'(\phi_i)\}^{-2}$, $i = 1,\ldots,n$. The joint iterative process for obtaining the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ takes the form

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^\top\boldsymbol{\Phi}^{(m)}\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Phi}^{(m)}\mathbf{W}^{(m)}\mathbf{y}^{*(m)} \tag{3}$$

and

$$\boldsymbol{\gamma}^{(m+1)} = (\mathbf{Z}^\top\mathbf{P}^{(m)}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{P}^{(m)}\mathbf{z}^{*(m)}, \tag{4}$$

for $m = 0,1,2,\ldots$, where $\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{-1/2}\mathbf{V}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})$ and $\mathbf{z}^* = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{V}_\gamma^{-1}\mathbf{H}_\gamma(\mathbf{t}-\boldsymbol{\mu}_T)$ are the modified dependent variables and $\mathbf{V}_\gamma = \mathrm{diag}\{-d''(\phi_1),\ldots,-d''(\phi_n)\}$. Note that $\mathbf{P} = \mathbf{V}_\gamma\mathbf{H}_\gamma^{-2}$. This joint iterative process is solved by alternating Eqs. (3)–(4) until convergence. Starting values may be the maximum likelihood estimates (MLEs) from the generalized linear model (GLM) with constant dispersion. The iterative process for obtaining the REMLEs takes the same form as (3)–(4) with the quantities $\mathbf{P}$ and $\mathbf{z}^*$ being modified appropriately (see, for instance, Smyth and Verbyla, 1999). Fahrmeir and Tutz (2001) presented some regularity conditions for attaining the asymptotic normality of the parameter estimates in GLMs. Assuming that such regularity conditions are extended for DGLMs, one has for large n that $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta},\mathbf{K}_{\beta\beta}^{-1})$ and $\hat{\boldsymbol{\gamma}} \sim \mathrm{N}_q(\boldsymbol{\gamma},\mathbf{K}_{\gamma\gamma}^{-1})$. Due to the orthogonality between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, one has asymptotic independence between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$. DGLMs may be performed by using, for instance, the packages `dglm` and `gamlss` available in `R` software.

3. Diagnostic methods

3.1. Leverage

The main idea behind the concept of leverage is that of evaluating the influence of each response on its own predicted value. In DGLMs, the influence of $\mathbf{y}$ on $\hat{\mathbf{y}}$ and $\mathbf{t}$ on $\hat{\mathbf{t}}$ may be well represented by the principal diagonal elements of the $n\times n$ matrices $(\partial\hat{\mathbf{y}}/\partial\mathbf{y}^\top)$ and $(\partial\hat{\mathbf{t}}/\partial\mathbf{t}^\top)$, respectively. Using results from Wei et al. (1998), we find the generalized leverage