Comput Stat (2013) 28:1639–1661
DOI 10.1007/s00180-012-0370-9

ORIGINAL PAPER

Multiple deletion diagnostics in beta regression models

Li-Chu Chien

Received: 9 October 2011 / Accepted: 8 September 2012 / Published online: 23 October 2012

Abstract We consider the problem of identifying multiple outliers in a general class of beta regression models proposed by Ferrari and Cribari-Neto (J Appl Stat 31:799–815, 2004). The currently available single-case deletion diagnostic measures, e.g., the standardized weighted residual (SWR), the Cook-like distance (LD), etc., often fail to identify multiple outlying observations, because they suffer from the well-known problems of masking and swamping effects. In this article, we develop group deletion diagnostic measures, such as generalized SWR, generalized LD, generalized DFFITS and generalized DFBETAS, and suggest a simple procedure for identifying multiple outliers using these. The performance of the proposed methods is investigated through simulation studies and two practical examples.

Keywords Beta regression · Multiple outliers · Generalized SWR · Generalized LD · Generalized DFFITS · Generalized DFBETAS

1 Introduction

Many fields of studies involve data in the form of percentages, rates or proportions that are measured continuously in the open interval (0, 1). For example, one may be interested in modeling the proportion of income spent on food as a function of the level of income and the number of persons in the household. The beta distribution is a flexible and useful tool for modeling data on the standard unit interval (0, 1), since the beta density can display quite different shapes depending on the values of the parameters that index the distribution; see, for example, Kieschnick and McCullough (2003).

L.-C. Chien (✉)
Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan
e-mail: lcchien@nctu.edu.tw

Ferrari and Cribari-Neto (2004) proposed a class of beta regression models which is the class of models derived from generalized linear models (GLMs), except the response variable is not from a linear exponential family distribution. Under generalized beta linear models (GBLMs), they developed maximum likelihood inference including parameter and interval estimation, and also hypothesis tests. They provided complete inference tools for the new class of models. The tools are freely available in the R package *betareg*. See Cribari-Neto and Zeileis (2010) for details. Hence, the execution of the beta regression techniques in practical problems is convenient.

Identifying observations that may affect the results of a regression analysis is a fundamental step in regression model building processes. In general, observations that lie well outside the majority of the data are termed outliers, in the sense that outliers come from a different probability distribution or from a different deterministic model than the mass of the data. The existence of outliers always distorts the outcome and accuracy of regression results, and hence, outliers must be detected in the regression analysis processes.

In GBLMs, to identify the outlying observations that depart from the postulated model of the bulk of the data, some authors have provided guidelines for diagnostic analysis. For example, Ferrari and Cribari-Neto (2004) provided some diagnostic measures to identify atypical observations and to detect model misspecification. Espinheira et al. (2008a) proposed two new beta residuals and numerically compared their behavior to those originally suggested by Ferrari and Cribari-Neto (2004). The results indicate a preference for one of the new residuals, more specifically the residual that accounts for the different leverages of the observations. On the other hand, Espinheira et al. (2008b) developed the Cook-like distance (LD) to measure the effects of influential observations on regression parameter estimates of GBLMs.

These currently available outlier measures and influence diagnostics seem to be available only when the data merely contain a single outlier. However, if the data contain more than one outlying observation, these existing methods may become ineffective, due to the problems of masking and swamping effects. Hence, multiple outlier detection methods that are free from these problems are proposed in this article.

This article unfolds as follows. Section 2 contains a concise review of GLBMs proposed by Ferrari and Cribari-Neto (2004). In the next section, we briefly introduce some of the current diagnostic tools, e.g., the standardized weighted residual (SWR), LD, etc. We also define the GBLM versions of the influence measures DFFITS and DFBETAS in this section. In the section after, we introduce SWR, LD, DFFITS and DFBETAS based on group deletion techniques and suggest an easy procedure for detecting multiple outliers using these group deletion diagnostics. Sections 5 and 6 illustrate applications of these newly proposed deletion diagnostic methods in simulated and real data examples, respectively. Finally, in Sect. 7, some conclusions about these proposed diagnostic measures are set out.

2 Beta regression model

The probability function for a single response variable Y in a GBLM (Ferrari and Cribari-Neto 2004) is

$$f(y;\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\quad 0<y<1 \tag{1}$$

where $0<\mu<1$, $\phi>0$ and $\Gamma(\cdot)$ is the gamma function. The mean and variance of Y are $E(Y)=\mu$ and $Var(Y)=V(\mu)/(1+\phi)$, respectively, where $V(\mu)=\mu(1-\mu)$ is a variance function. Note that ϕ can be viewed as a precision parameter, in the sense that, for a fixed mean μ, the variance of Y decreases as ϕ increases. Here our concern is with the mean parameter μ, and ϕ may be viewed as a nuisance parameter.

Now consider observations $y_1,\ldots,y_n$ which are regarded as realizations of independent random variables $Y_1,\ldots,Y_n$ and each $y_i, i=1,\ldots,n$, follows the density (1) with mean μ_i and unknown precision ϕ. The mean of Y_i involves explanatory variables through a link function $g(\cdot)$, so that $g(\mu_i)=\eta_i$ where $\eta_i=\boldsymbol{x}_i^T\boldsymbol{\beta}=\beta_1x_{i,1}+\beta_2x_{i,2}+\cdots+\beta_px_{i,p}$. Here η_i is a linear predictor, $\boldsymbol{x}_i=(x_{i,1},\ldots,x_{i,p})^T$ is a p-vector of explanatory variables, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$ is a p-vector of unknown regression parameters, and $g(\cdot)$ is a strictly monotonic and twice differentiable link function that forms a mapping from the interval (0, 1) to $\mathbb{R}$. The parameters $\boldsymbol{\beta}$ and ϕ can be estimated by maximum likelihood methods that can easily be implemented through the R package *betareg*. For details, see Cribari-Neto and Zeileis (2010).

3 Measures of influence based on single-case deletion diagnostics

In this section, we succinctly review the currently available single-case deletion diagnostic methods. We also define the GBLM versions of the influence measures DFFITS and DFBETAS according to the classical versions of DFFITS and DFBETAS established under the normal regression settings and discussed in Subsection 2.1 of Belsley et al. (1980).

3.1 Assessing the influence of case deletion on residuals

This subsection is focused on the single-case outlier identification tools which use the residuals to detect the atypical observations and check model adequacy. We review the weighted residual (WR) and SWR proposed by Espinheira et al. (2008a).

WR and SWR. Espinheira et al. (2008a) defined WR for point i as

$$r_i^*=\frac{y_i^*-\widehat{\mu}_i^*}{\sqrt{\widehat{\phi}\widehat{v}_i}} \tag{2}$$

where $y_i^*=\log(y_i/(1-y_i))$, $\widehat{\mu}_i^*=\psi(\widehat{\mu}_i\widehat{\phi})-\psi((1-\widehat{\mu}_i)\widehat{\phi})$, $\widehat{v}_i=\{\psi'(\widehat{\mu}_i\widehat{\phi})+\psi'((1-\widehat{\mu}_i)\widehat{\phi})\}$ and $\psi'(\cdot)$ is the trigamma function. Here the symbol "$\widehat{\ }$" is used for quantities that are evaluated at parameter values of the maximum likelihood solution based on all observations. Then the standardized version, SWR, for point i is defined by

$$r_i^{ww}=\frac{r_i^*}{\sqrt{(1-\widehat{h}_i)/\widehat{\phi}}}=\frac{y_i^*-\widehat{\mu}_i^*}{\sqrt{\widehat{v}_i\,(1-\widehat{h}_i)}} \tag{3}$$

where $\widehat{h}_i$ is the ith diagonal element of $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{W}}^{1/2}$ and $\widehat{\boldsymbol{W}}^{1/2}$ is a symmetric square root of $\widehat{\boldsymbol{W}}$. Here $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $\widehat{\boldsymbol{W}} = \widehat{\phi}\widehat{\boldsymbol{G}}\widehat{\boldsymbol{V}}\widehat{\boldsymbol{G}}$ with diagonal matrices $\widehat{\boldsymbol{G}} = Diag(1/g'(\widehat{\mu}_1), \ldots, 1/g'(\widehat{\mu}_n))$ and $\widehat{\boldsymbol{V}} = Diag(\widehat{v}_1, \ldots, \widehat{v}_n)$. An excessively large or small value of WR_i or SWR_i indicates point i having an unusual residual.

3.2 Assessing the influence of case deletion on regression parameter estimates

This subsection is concerned with the single-case influence identification tools which measure the influence of deleting one observation on the regression parameter estimates or on the fitted values. We review LD proposed by Espinheira et al. (2008b) and propose the GBLM versions of the influence measures DFFITS and DFBETAS.

LD. Espinheira et al. (2008b) defined LD for point i as

$$\mathrm{LD}_i = \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(-i)}\right)^T \left(\widehat{\phi}\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X}\right)\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(-i)}\right)$$

where $\widehat{\boldsymbol{\beta}}^{(-i)}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ without the ith observation. From the approximate relation

$$\widehat{\boldsymbol{\beta}}^{(-i)} \approx \widehat{\boldsymbol{\beta}} - \frac{\left(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\widehat{w}_i^{1/2}}{1-\widehat{h}_i}r_i^* \tag{4}$$

where $\widehat{w}_i^{1/2}$ is the ith diagonal entry of $\widehat{\boldsymbol{W}}^{1/2}$, it follows that

$$\mathrm{LD}_i \approx \left(\frac{r_i^*}{1-\widehat{h}_i}\right)^2 \widehat{\phi}\widehat{h}_i = \left(r_i^{ww}\right)^2 \frac{\widehat{h}_i}{1-\widehat{h}_i}. \tag{5}$$

A large LD_i indicates that point i has a large impact on $\widehat{\boldsymbol{\beta}}$.

DFFITS. We define DFFITS for point i as

$$\mathrm{DFFITS}_i = \frac{\widehat{w}_i^{1/2}\boldsymbol{x}_i^T\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(-i)}\right)}{\sqrt{\widehat{h}_i/\widehat{\phi}^{(-i)}}}$$

which, in view of (4), is expressed by

$$\mathrm{DFFITS}_i \approx \frac{r_i^*}{\sqrt{1-\widehat{h}_i}}\frac{\sqrt{\widehat{h}_i\widehat{\phi}^{(-i)}}}{\sqrt{1-\widehat{h}_i}} = r_i^{ww}\sqrt{\frac{\widehat{\phi}^{(-i)}}{\widehat{\phi}}}\sqrt{\frac{\widehat{h}_i}{1-\widehat{h}_i}} \tag{6}$$

where $\widehat{\phi}^{(-i)}$ is the MLE of ϕ with observation i omitted. Using (6), we obtain, from (5), that

$$\mathrm{LD}_i = (\mathrm{DFFITS}_i)^2 \frac{\widehat{\phi}}{\widehat{\phi}^{(-i)}}.$$

An extreme DFFITS_i implies that observation i is influential on the weighted fit, $\widehat{w}_i^{1/2}\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$.

DFBETAS. We define DFBETAS for point i for β_j as

$$\mathrm{DFBETAS}_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}_j^{(-i)}}{\sqrt{\left(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X}\right)_j^{-1}/\widehat{\phi}^{(-i)}}}$$

where $(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})_j^{-1}$ is the jth diagonal element of the inverse of $\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X}$, $(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$. Let c_{ji} be the jth component of the p-vector $(\boldsymbol{X}^T\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{w}_i^{1/2}$. We then have

$$\mathrm{DFBETAS}_{ij} \approx r_i^* \frac{\sqrt{\widehat{\phi}^{(-i)}}}{1-\widehat{h}_i} \frac{c_{ji}}{\sqrt{\sum_{i=1}^n c_{ji}^2}} = r_i^{ww}\sqrt{\frac{\widehat{\phi}^{(-i)}}{\widehat{\phi}(1-\widehat{h}_i)}} \frac{c_{ji}}{\sqrt{\sum_{i=1}^n c_{ji}^2}}. \tag{7}$$

A large absolute value of $\mathrm{DFBETAS}_{ij}$ shows that observation i is influential on $\widehat{\beta}_j$.

4 Measures of influence based on group deletion diagnostics

In this section, we introduce the group deletion diagnostic measures and suggest a diagnostic procedure for detecting multiple outlying observations using these. We assume that d observations among a set of n observations are unusual observations and omitted before the fitting of the model. Let $[\boldsymbol{R}]$ index a set of the $(n-d)$ observations that are remaining in the analysis after deleting a set of the d unusual observations indexed by $\boldsymbol{D}$.

Let $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^T$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ and $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)^T$ be the $n \times 1$ vectors. Without loss of generality, we assume that observations to be deleted are the last d components of $\boldsymbol{y}^*$ and $\boldsymbol{X}$, so that $\boldsymbol{y}^{*T} = (\boldsymbol{y}_{[\boldsymbol{R}]}^{*T}, \boldsymbol{y}_{\boldsymbol{D}}^{*T})$ and $\boldsymbol{X}^T = (\boldsymbol{X}_{[\boldsymbol{R}]}^T, \boldsymbol{X}_{\boldsymbol{D}}^T)$. Let $\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]}, \widehat{\phi}_{[\boldsymbol{R}]}, \widehat{\boldsymbol{\mu}}_{[\boldsymbol{R}]}, \widehat{\boldsymbol{\mu}}_{[\boldsymbol{R}]}^*, \widehat{\boldsymbol{V}}_{[\boldsymbol{R}]}, \widehat{\boldsymbol{G}}_{[\boldsymbol{R}]}$ and $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}$ be the MLEs, respectively, of $\boldsymbol{\beta}, \phi, \boldsymbol{\mu}, \boldsymbol{\mu}^*, \boldsymbol{V}, \boldsymbol{G}$ and $\boldsymbol{W}$ without the d observations in the deletion set $\boldsymbol{D}$. Let $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}^{-1}$ be the inverse of $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}$. Then write

$$z_{[\boldsymbol{R}]} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]} + \widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}^{-1}\widehat{\boldsymbol{G}}_{[\boldsymbol{R}]}\left(\boldsymbol{y}^* - \widehat{\boldsymbol{\mu}}_{[\boldsymbol{R}]}^*\right) \tag{8}$$

and partition $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}$ as

$$\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]} = \widehat{\phi}_{[\boldsymbol{R}]}\widehat{\boldsymbol{G}}_{[\boldsymbol{R}]}\widehat{\boldsymbol{V}}_{[\boldsymbol{R}]}\widehat{\boldsymbol{G}}_{[\boldsymbol{R}]} = \begin{bmatrix} \widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]} & \boldsymbol{0} \\ \boldsymbol{0}^T & \widehat{\boldsymbol{W}}_{\boldsymbol{D}[\boldsymbol{R}]} \end{bmatrix}$$

with diagonal matrices $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}$ and $\widehat{\boldsymbol{W}}_{\boldsymbol{D}[\boldsymbol{R}]}$ being related to observations from the $[\boldsymbol{R}]$ and $\boldsymbol{D}$ sets, respectively, in which $\boldsymbol{0}$ is a zero matrix of order $(n-d)\times d$. When a group of observations $\boldsymbol{D}$ is omitted, we define the leverage measure for case i as $\widehat{h}_{i[\boldsymbol{R}]} = \widehat{w}_{i[\boldsymbol{R}]}^{1/2}\boldsymbol{x}_i^T(\boldsymbol{X}_{[\boldsymbol{R}]}^T\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]})^{-1}\boldsymbol{x}_i\widehat{w}_{i[\boldsymbol{R}]}^{1/2}$ where $\widehat{w}_{i[\boldsymbol{R}]}^{1/2}$ is the ith element of $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}^{1/2}$ that is a symmetric square root of $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}]}$.

4.1 Identifying multiple outliers using group-deleted versions of WR and SWR

The generalized WR (GWR) and the generalized SWR (GSWR), based on group deletion techniques, are proposed by the single-case outlier detection measures WR and SWR.

GWR. For point i from the $[\boldsymbol{R}]$ set, WR is defined, in view of (2), as

$$r_{i[\boldsymbol{R}]}^{*} = \frac{y_i^{*} - \widehat{\mu}_{i[\boldsymbol{R}]}^{*}}{\sqrt{\widehat{\phi}_{[\boldsymbol{R}]}\widehat{v}_{i[\boldsymbol{R}]}}} \tag{9}$$

where $\widehat{\mu}_{i[\boldsymbol{R}]}^{*} = \psi(\widehat{\mu}_{i[\boldsymbol{R}]}\widehat{\phi}_{[\boldsymbol{R}]}) - \psi((1-\widehat{\mu}_{i[\boldsymbol{R}]})\widehat{\phi}_{[\boldsymbol{R}]})$ and $\widehat{v}_{i[\boldsymbol{R}]}$ is the ith diagonal unit of $\widehat{\boldsymbol{V}}_{[\boldsymbol{R}]}$. Using (8), $r_{i[\boldsymbol{R}]}^{*}$ is expressed in the alternative form by

$$r_{i[\boldsymbol{R}]}^{*} = \widehat{w}_{i[\boldsymbol{R}]}^{1/2}\left(z_{i[\boldsymbol{R}]} - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]}\right) \tag{10}$$

which is called the internal WR (IWR), because the point i is from the $[\boldsymbol{R}]$ set.

In view of (10), WR for point i from the $\boldsymbol{D}$ set is defined as $r_{i[\boldsymbol{R}+i]}^{*} = \widehat{w}_{i[\boldsymbol{R}]}^{1/2}(z_{i[\boldsymbol{R}]} - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}+i]})$ where notations "$\widehat{\ }$" and "$[\boldsymbol{R}+i]$" are used for quantities that are evaluated at parameter values of the maximum likelihood solution based on the remaining set $[\boldsymbol{R}+i]$, which consists of $(n-d+1)$ observations with the point i from the $\boldsymbol{D}$ set and the others from the $[\boldsymbol{R}]$ set. Using

$$\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}+i]} \approx \widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]} + \frac{\left(\boldsymbol{X}_{[\boldsymbol{R}]}^T\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]}\right)^{-1}\boldsymbol{x}_i\widehat{w}_{i[\boldsymbol{R}]}^{1/2}}{1+\widehat{h}_{i[\boldsymbol{R}]}}r_{i[\boldsymbol{R}]}^{*}, \tag{11}$$

$r_{i[\boldsymbol{R}+i]}^{*}$ is equivalently expressed in the form

$$r_{i[\boldsymbol{R}+i]}^{*} = \frac{r_{i[\boldsymbol{R}]}^{*}}{1+\widehat{h}_{i[\boldsymbol{R}]}} \tag{12}$$

which is called the external WR (EWR), because the point i is from the $\boldsymbol{D}$ set. Then GWR_i for any data points is defined by (9) for i in $[\boldsymbol{R}]$ and by (12) for i in $\boldsymbol{D}$.

GSWR. For point i from the $[\boldsymbol{R}]$ set, the internal SWR (ISWR) is defined, using (3), as

$$Ir_{i[\boldsymbol{R}]}^{ww} = \frac{r_{i[\boldsymbol{R}]}^{*}}{\sqrt{\left(1-\widehat{h}_{i[\boldsymbol{R}]}\right)/\widehat{\phi}_{[\boldsymbol{R}]}}}. \tag{13}$$

On the other hand, for point i from the $\boldsymbol{D}$ set, the external SWR (ESWR) is defined, using (13), as $Er^{ww}_{i[\boldsymbol{R}+i]} = r^{*}_{i[\boldsymbol{R}+i]}/\sqrt{(1-\widehat{h}_{i[\boldsymbol{R}+i]})/\widehat{\phi}_{[\boldsymbol{R}+i]}}$. By writing

$$\widehat{h}_{i[\boldsymbol{R}+i]} = \widehat{w}^{1/2}_{i[\boldsymbol{R}]}\boldsymbol{x}_i^T\left(\boldsymbol{X}^T_{[\boldsymbol{R}]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]} + \boldsymbol{x}_i\widehat{w}_{i[\boldsymbol{R}]}\boldsymbol{x}_i^T\right)^{-1}\boldsymbol{x}_i\widehat{w}^{1/2}_{i[\boldsymbol{R}]} = \frac{\widehat{h}_{i[\boldsymbol{R}]}}{1+\widehat{h}_{i[\boldsymbol{R}]}} \quad (14)$$

and using (12), ESWR is seen to be equal to

$$Er^{ww}_{i[\boldsymbol{R}+i]} = \frac{r^{*}_{i[\boldsymbol{R}]}}{\sqrt{\left(1+\widehat{h}_{i[\boldsymbol{R}]}\right)/\widehat{\phi}_{[\boldsymbol{R}+i]}}}. \quad (15)$$

Then GSWR_i for any data point is defined by (13) for i in $[\boldsymbol{R}]$ and by (15) for i in $\boldsymbol{D}$.

In fact, group deletion residuals for both the $[\boldsymbol{R}]$ set and the $\boldsymbol{D}$ set, which are measured based on a similar scale, have received a great deal of attention in the literature in recent years. For example, similar residuals were derived by Hadi and Simonoff (1993) for detecting multiple outliers in linear models and, later, a similar diagnostic technique was sketched by Atkinson (1994). In addition, this type of residual was introduced, under linear regression and logistic regression, by Imon (2005) and Imon and Hadi (2008).

4.2 Identifying multiple outliers using group-deleted versions of LD, DFFITS and DFBETAS

The generalized LD (GLD), the generalized DFFITS (GDFFITS) and the generalized DFBETAS (GDFBETAS), based on group deletion techniques, are proposed by the single-case influence detection measures LD, DFFITS and DFBETAS.

GLD. For point i from the $[\boldsymbol{R}]$ set, the internal LD is computed by

$$\mathrm{ILD}_i = \left(\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]} - \widehat{\boldsymbol{\beta}}^{(-i)}_{[\boldsymbol{R}]}\right)^T\left(\widehat{\phi}_{[\boldsymbol{R}]}\boldsymbol{X}^T_{[\boldsymbol{R}]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]}\right)\left(\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]} - \widehat{\boldsymbol{\beta}}^{(-i)}_{[\boldsymbol{R}]}\right)$$

whereas, for point i from the $\boldsymbol{D}$ set, the external LD is computed by

$$\mathrm{ELD}_i = \left(\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}+i]} - \widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]}\right)^T\left(\widehat{\phi}_{[\boldsymbol{R}+i]}\boldsymbol{X}^T_{[\boldsymbol{R}+i]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}+i][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}+i]}\right)\left(\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}+i]} - \widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]}\right)$$

where $\widehat{\boldsymbol{W}}_{[\boldsymbol{R}+i][\boldsymbol{R}]}$ is an $(n-d+1)\times(n-d+1)$ diagonal matrix, in which the first $(n-d)$ diagonal units are relative to observations from the $[\boldsymbol{R}]$ set and the last diagonal unit is relative to the point i from the $\boldsymbol{D}$ set. GLD_i is derived by combining ILD_i and ELD_i. Through (11) and using $\boldsymbol{X}^T_{[\boldsymbol{R}+i]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}+i][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}+i]} = \boldsymbol{X}^T_{[\boldsymbol{R}]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]} + \boldsymbol{x}_i\widehat{w}_{i[\boldsymbol{R}]}\boldsymbol{x}_i^T$ and

$$\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]} \approx \widehat{\boldsymbol{\beta}}^{(-i)}_{[\boldsymbol{R}]} + \frac{\left(\boldsymbol{X}^T_{[\boldsymbol{R}]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]}\right)^{-1}\boldsymbol{x}_i\widehat{w}^{1/2}_{i[\boldsymbol{R}]}}{1-\widehat{h}_{i[\boldsymbol{R}]}}r^{*}_{i[\boldsymbol{R}]}, \quad (16)$$

GLD_i is expressed in terms of $Ir^{ww}_{i[\boldsymbol{R}]}$ and $\widehat{h}_{i[\boldsymbol{R}]}$ as

$$\mathrm{GLD}_i = \begin{cases} (Ir^{ww}_{i[\boldsymbol{R}]})^2 \frac{\widehat{h}_{i[\boldsymbol{R}]}}{1-\widehat{h}_{i[\boldsymbol{R}]}} & \text{if } i \in [\boldsymbol{R}] \\ \left(Ir^{ww}_{i[\boldsymbol{R}]}\right)^2 \frac{(1-\widehat{h}_{i[\boldsymbol{R}]})\widehat{h}_{i[\boldsymbol{R}]}}{1+\widehat{h}_{i[\boldsymbol{R}]}} \frac{\widehat{\phi}_{[\boldsymbol{R}+i]}}{\widehat{\phi}_{[\boldsymbol{R}]}} & \text{if } i \notin [\boldsymbol{R}]. \end{cases}$$

GDFFITS. The internal DFFITS (IDFFITS), for point i from the $[\boldsymbol{R}]$ set, is defined by

$$\mathrm{IDFFITS}_i = \frac{\widehat{w}^{1/2}_{i[\boldsymbol{R}]} \boldsymbol{x}_i^T \left(\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]} - \widehat{\boldsymbol{\beta}}^{(-i)}_{[\boldsymbol{R}]}\right)}{\sqrt{\widehat{h}_{i[\boldsymbol{R}]}/\widehat{\phi}^{(-i)}_{[\boldsymbol{R}]}}}$$

whereas the external DFFITS (EDFFITS), for point i from the $\boldsymbol{D}$ set, is defined by

$$\mathrm{EDFFITS}_i = \frac{\widehat{w}^{1/2}_{i[\boldsymbol{R}]} \boldsymbol{x}_i^T \left(\widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}+i]} - \widehat{\boldsymbol{\beta}}_{[\boldsymbol{R}]}\right)}{\sqrt{\widehat{h}_{i[\boldsymbol{R}+i]}/\widehat{\phi}_{[\boldsymbol{R}]}}}.$$

Together these $\mathrm{IDFFITS}_i$ and $\mathrm{EDFFITS}_i$ give $\mathrm{GDFFITS}_i$ for point i in the entire data set. Using (16) in conjunction with (11) and (14), it turns out that $\mathrm{GDFFITS}_i$ is displayed in terms of $Ir^{ww}_{i[\boldsymbol{R}]}$ and $\widehat{h}_{i[\boldsymbol{R}]}$ as

$$\mathrm{GDFFITS}_i = \begin{cases} Ir^{ww}_{i[\boldsymbol{R}]} \frac{\sqrt{\widehat{\phi}^{(-i)}_{[\boldsymbol{R}]}\widehat{h}_{i[\boldsymbol{R}]}}}{\sqrt{\widehat{\phi}_{[\boldsymbol{R}]}(1-\widehat{h}_{i[\boldsymbol{R}]})}} & \text{if } i \in [\boldsymbol{R}] \\ Ir^{ww}_{i[\boldsymbol{R}]} \frac{\sqrt{(1-\widehat{h}_{i[\boldsymbol{R}]})\widehat{h}_{i[\boldsymbol{R}]}}}{\sqrt{1+\widehat{h}_{i[\boldsymbol{R}]}}} & \text{if } i \notin [\boldsymbol{R}]. \end{cases}$$

After some algebraic manipulation, it is observed

$$\mathrm{GLD}_i = \begin{cases} (\mathrm{GDFFITS}_i)^2 \frac{\widehat{\phi}_{[\boldsymbol{R}]}}{\widehat{\phi}^{(-i)}_{[\boldsymbol{R}]}} & \text{if } i \in [\boldsymbol{R}] \\ (\mathrm{GDFFITS}_i)^2 \frac{\widehat{\phi}_{[\boldsymbol{R}+i]}}{\widehat{\phi}_{[\boldsymbol{R}]}} & \text{if } i \notin [\boldsymbol{R}]. \end{cases}$$

GDFBETAS. For point i from the $[\boldsymbol{R}]$ set, we define the internal DFBETAS (IDFBETAS), for β_j, as

$$\mathrm{IDFBETAS}_{ij} = \frac{\widehat{\beta}_{j[\boldsymbol{R}]} - \widehat{\beta}^{(-i)}_{j[\boldsymbol{R}]}}{\sqrt{\left(\boldsymbol{X}^T_{[\boldsymbol{R}]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]}\right)^{-1}_j / \widehat{\phi}^{(-i)}_{[\boldsymbol{R}]}}}.$$

For point i from the $\boldsymbol{D}$ set, we define the external DFBETAS (EDFBETAS), for β_j, as

$$\mathrm{EDFBETAS}_{ij} = \frac{\widehat{\beta}_{j[\boldsymbol{R}+i]} - \widehat{\beta}_{j[\boldsymbol{R}]}}{\sqrt{\left(\boldsymbol{X}^T_{[\boldsymbol{R}+i]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}+i][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}+i]}\right)^{-1}_j \Big/ \widehat{\phi}_{[\boldsymbol{R}]}}}.$$

Collecting these IDFBETAS$_{ij}$ and EDFBETAS$_{ij}$, for any point i in the data set, GDFBETAS, for β_j, is exhibited in terms of $Ir^{ww}_{i[\boldsymbol{R}]}$ and $\widehat{h}_{i[\boldsymbol{R}]}$ as

$$\text{GDFBETAS}_{ij} = \begin{cases} Ir^{ww}_{i[\boldsymbol{R}]} \dfrac{\sqrt{\widehat{\phi}^{(-i)}_{[\boldsymbol{R}]}}}{\sqrt{\widehat{\phi}_{[\boldsymbol{R}]}(1-\widehat{h}_{i[\boldsymbol{R}]})}} \dfrac{c_{ji[\boldsymbol{R}]}}{\sqrt{\sum_{i\in\boldsymbol{R}} c^2_{ji[\boldsymbol{R}]}}} & \text{if } i \in [\boldsymbol{R}] \\ Ir^{ww}_{i[\boldsymbol{R}]} \dfrac{\sqrt{1-\widehat{h}_{i[\boldsymbol{R}]}}}{1+\widehat{h}_{i[\boldsymbol{R}]}} \dfrac{c_{ji[\boldsymbol{R}]}}{\sqrt{\sum_{i\in\boldsymbol{R}} c^2_{ji[\boldsymbol{R}]} - \left(c^2_{ji[\boldsymbol{R}]}/(1+\widehat{h}_{i[\boldsymbol{R}]})\right)}} & \text{if } i \notin [\boldsymbol{R}] \end{cases}$$

where $c_{ji[\boldsymbol{R}]}$ is the jth component of the p-vector $(\boldsymbol{X}^T_{[\boldsymbol{R}]}\widehat{\boldsymbol{W}}_{[\boldsymbol{R}][\boldsymbol{R}]}\boldsymbol{X}_{[\boldsymbol{R}]})^{-1}\boldsymbol{x}_i\,\widehat{w}^{1/2}_{i[\boldsymbol{R}]}$.

Similar diagnostic measures are derived by Imon (2005) who proposed the group deletion versions of LD and DFFITS for the identification of multiple influential observations in linear regression. Nurunnabi et al. (2010) developed the generalized DFFITS in logistic regression for the same purpose.

4.3 The diagnostic algorithm

In this subsection, attention is given to the introduction of a test procedure for the detection of multiple outliers through the group deletion diagnostic measures, GSWR$_i$, GLD$_i$, GDFFITS$_i$ and GDFBETAS$_i$, respectively. The main idea of the proposed method is to first form a basic subset of one fourth of the data which is possibly free from potential outliers and then employ the group deletion diagnostics, based on the basic subset, in identifying outlying observations. The detailed diagnostic algorithm is illustrated with GSWR$_i$ as follows.

Step 0 Fit the regression model to the full data and compute r^{ww}_i for $i = 1, \ldots, n$.
Step 1 Arrange the n observations in ascending order according to the absolute values of r^{ww}_i, $|r^{ww}_i|$, $i = 1, \ldots, n$. Then the first $\{n/4\}$ observations form the original $[\boldsymbol{R}]$ set, where $\{n/4\}$ is the integer part of $n/4$ and represents the initial size of the $[\boldsymbol{R}]$ set. If the initial subset is not of full rank, increase the initial subset by as many observations as needed for the initial subset to become full rank (the observations are added according to their ranked order).
Step 2 Fit a regression model to the current $[\boldsymbol{R}]$ set and compute GSWR$_i$.
Step 3 Arrange observations in accordance with an increasing order of $|\text{GSWR}_i|$ and let GSWR$_{(s+1)}$ be the $(s+1)$th order statistic of $|\text{GSWR}_i|$, where s is the current size of the $[\boldsymbol{R}]$ set.

(a) If $\text{GSWR}_{(s+1)} > 2c(s)$ with $c(s) = \sqrt{\widehat{\phi}_{[\boldsymbol{R}]}/\widehat{\phi}}\sqrt{\sum_{i=1}^n |\text{GSWR}_i|/\sum_{i=1}^n |\text{SWR}_i|}$, then declare all members satisfying $|\text{GSWR}_i| > 2c(s)$ as outliers, and stop.
(b) Otherwise, the current $[\boldsymbol{R}]$ set is replaced by the first $(s+1)$ ordered observations. If $s + 1 = n$, then go to Step 4; otherwise return to Step 2.

Step 4 GSWR$_i$ is recalculated with all data points involved into the $[\boldsymbol{R}]$ set. Then declare all observations satisfying $|\text{GSWR}_i| > 2$ as outliers and stop.

Here we suggest to use the cut-off values of 2 and $2c(s)$ for $|\text{SWR}_i|$ and $|\text{GSWR}_i|$, respectively. That is because, SWR$_i$ and GSWR$_i$ are standardized quantities. Hence there will be very few SWR$_i$'s larger than 2 and GSWR$_i$'s larger than $2c(s)$, in which

$c(s)$ plays the role as an adjustment, which adjusts the effect from the differences between the sample size, n, and the current size of the $[\boldsymbol{R}]$ set, s. It should be explained that, in practice, the criteria of $|\mathrm{SWR}_i|$ and $|\mathrm{GSWR}_i|$ can be flexibly defined by other cut-off values, e.g., 2.5 and $2.5c(s)$, 3 and $3c(s)$, etc.

Note that if there are a few outliers involved in the initially basic subset of size $\{n/4\}$, these outliers are still revealed as outlying because these observations should be gradually excluded from the subsets with sizes $\{n/4\}+1$, $\{n/4\}+2, \ldots$, respectively.

The diagnostic algorithms for GLD_i, $\mathrm{GDFFITS}_i$ and $\mathrm{GDFBETAS}_{ij}$ are similar, except the cut-off values, and hence they are omitted. Their respective suggested cut-off values are as follows.

Under the assumption that the weighted matrix $\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}$ is of full rank, the average of $\widehat{h}_i$ is p/n, and then LD_i in (5) becomes

$$\mathrm{LD}_i = \left(r_i^{ww}\right)^2 \frac{p}{n-p}.$$

Hence we suggest to use the cut-off values of $4p/n$ and $4c^2(s)\,p/s$ for LD_i and GLD_i, respectively. Under the same assumption, DFFITS_i reduces from (6) to

$$\mathrm{DFFITS}_i = r_i^{ww}\sqrt{\frac{\widehat{\phi}^{(-i)}}{\widehat{\phi}}}\sqrt{\frac{p}{n-p}}.$$

Hence we suggest to use the cut-off values of $2\sqrt{p/n}$ and $2c(s)\sqrt{p/s}$ for $|\mathrm{DFFITS}_i|$ and $|\mathrm{GDFFITS}_i|$, respectively.

On the other hand, in the special case of location, i.e., in the special case with $\widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}$ that is an $n \times 1$ vector of ones, $\mathrm{DFBETAS}_{ij}$ reduces from (7) to

$$\mathrm{DFBETAS}_i = \frac{y_i^* - \widehat{\mu}_i^*}{\sqrt{\widehat{v}_i}}\sqrt{\frac{\widehat{\phi}^{(-i)}}{\widehat{\phi}}}\frac{\sqrt{n}}{n-1}$$

where the quantity $(y_i^* - \widehat{\mu}_i^*)/\sqrt{\widehat{v}_i}$ is a standardized quantity of y_i^*, since the expectation and variance of the random variable $Y_i^* = \log(Y_i/(1-Y_i))$ are $E(Y_i^*) = \mu_i^*$ and $Var(Y_i^*) = v_i$, respectively. Thus, the cut-off values of $\mathrm{DFBETAS}_{ij}$ and $\mathrm{GDFBETAS}_{ij}$ are suggested as $2/\sqrt{n}$ and $2c(s)/\sqrt{s}$, respectively.

Here it should be pointed out that the proposed diagnostic method has a resemblance to the forward search (FS) algorithm that was introduced by Hadi and Simonoff (1993) and Atkinson (1994) for identifying multiple outliers in linear models and in multivariate data, respectively. The proposed method and the FS algorithm have two similar stages. In the first stage, it is attempted to form a basic subset that is presumably free from potential outliers. In the second stage, it uses an appropriate diagnostic measure such as the adjusted residual or Cook distance to examine the potential outliers to see how extreme they are related to the basic subset. The possible outliers are then declared as outliers if they are greatly inconsistent with the majority of the data. Such a FS diagnostic technique was also applied to GLMs. For details, see Atkinson and Riani (2000, Chapter 6). A recent survey of theoretical development in work on the FS diagnostic was given in Atkinson et al. (2010), who tried to get suitable envelopes

for the outlier tests as the size of the current [$\boldsymbol{R}$] set grows. The modified FS diagnostic procedure has not been applied to beta regression models.

5 Simulations

In this section, simulation studies are carried out to investigate the finite sample performance of the proposed group deletion diagnostic measures. We consider the two beta models

$$\text{Model 1}: \ \text{logit}(\mu_i) = \log\left(\mu_i/(1-\mu_i)\right) = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

and

$$\text{Model 2}: \ \Phi\left(\mu_i\right) = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

with the logit and probit link functions, respectively. To show the different performance characters between GSWR_i, GLD_i, GDFFITS_i and GDFBETAS_i, numerical studies are performed for sample sizes $n = 50$ and 150 with the first $0.2n$ observations generated as outlying observations that have unusual residuals, the next $0.1n$ observations generated as outlying data points that are influential on $\widehat{\beta}_2$, and subsequently the next $0.1n$ observations generated as outlying data points that have an influence on $\widehat{\beta}_3$. To further contrast the differences between the performances based on the single-case and group deletion diagnostics, we consider two different outlier scenarios under each model. These scenarios are chosen because they are situations in which groups of outliers are influential for regression analyses but they are not easily identified by the single-case deletion diagnostics.

Let $\text{logit}^{-1}(\cdot)$ and $\Phi^{-1}(\cdot)$ be the inverses of $\text{logit}(\cdot)$ and $\Phi(\cdot)$, respectively. Now consider Model 1 with $n = 50$ and $\phi = 30$. Let $\beta_1 = \beta_2 = \beta_3 = -0.01$ and $\eta_i = -0.01 - 0.01x_{i,2} - 0.01x_{i,3}$. Then the usual observations $y_{21}, \ldots, y_{50}$ are independently generated from the beta distribution, $Beta(\mu_i\phi, (1-\mu_i)\phi)$, with $\mu_i = \text{logit}^{-1}(\eta_i)$ in which $x_{i,2}$ and $x_{i,3}$ are independently from the uniform distribution *Uniform* (0.85, 1.15). Under the outlier scenario (A), the outlying observations $y_1, \ldots, y_{10}$ that have extra small residuals are independently generated as the usual data points, except their means set by $\mu_i = \text{logit}^{-1}(\eta_i) - 0.35$. The outlying data points $y_{11}, \ldots, y_{15}$ that are influential on $\widehat{\beta}_2$ are independently from the beta distribution with $\mu_i = \text{logit}^{-1}(-0.01 + 0.85x_{i,2} - 0.01x_{i,3})$ where $x_{i,2}$ and $x_{i,3}$ are independently from *Uniform* (1.3, 1.4) and *Uniform* (0.85, 1.15). The outlying data points $y_{16}, \ldots, y_{20}$ that have an influence on $\widehat{\beta}_3$ are independently generated by the same processes, except their means set by $\mu_i = \text{logit}^{-1}(-0.01 - 0.01x_{i,2} + 0.85x_{i,3})$ with $x_{i,2}$ and $x_{i,3}$ respectively from *Uniform* (0.85, 1.15) and *Uniform* (1.3, 1.4). On the other hand, under the outlier scenario (B), the outliers $y_1, \ldots, y_{20}$ are generated as under the outlier scenario (A), except their respective corresponding means reset by $\mu_i = \text{logit}^{-1}(\eta_i) + 0.41$, $\mu_i = \text{logit}^{-1}(-0.01 - 0.6x_{i,2} - 0.01x_{i,3})$ and $\mu_i = \text{logit}^{-1}(-0.01 - 0.01x_{i,2} - 0.7x_{i,3})$ for observation i, $i = 1, \ldots, 10$, $i = 11, \ldots, 15$ and $i = 16, \ldots, 20$, respectively. Data for $n = 150$ are similarly generated by the same steps.

Next consider Model 2 with $n = 50$ and $\phi = 30$. The usual observations $y_{21}, \ldots, y_{50}$ are independently from $Beta(\mu_i\phi, (1-\mu_i)\phi)$ with $\mu_i = \Phi^{-1}(\eta_i)$

where $x_{i,2}$ and $x_{i,3}$ are independently from *Uniform* (0.85, 1.15). Under the outlier scenario (C), the outlying observations $y_1, \ldots, y_{10}$ that have extremely large residuals are independently generated as the usual data points, except their means set by $\mu_i = \Phi^{-1}(\eta_i) + 0.38$. The outlying data points $y_{11}, \ldots, y_{15}$ and $y_{16}, \ldots, y_{20}$ that respectively have an effect on $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are independently from the beta distribution with $\mu_i = \Phi^{-1}(-0.01-0.7x_{i,2}-0.01x_{i,3})$ and $\mu_i = \Phi^{-1}(-0.01-0.01x_{i,2}-0.7x_{i,3})$, respectively, where $x_{i,2}$ and $x_{i,3}$ are generated as in Model 1. Under the outlier scenario (D), the outliers $y_1, \ldots, y_{20}$ are generated as under the outlier scenario (C), except their respective corresponding means reset by $\mu_i = \Phi^{-1}(\eta_i) - 0.4$, $\mu_i = \Phi^{-1}(-0.01+0.4x_{i,2}-0.01x_{i,3})$ and $\mu_i = \Phi^{-1}(-0.01-0.01x_{i,2}+0.5x_{i,3})$, respectively. Data for $n = 150$ are also generated in the same way.

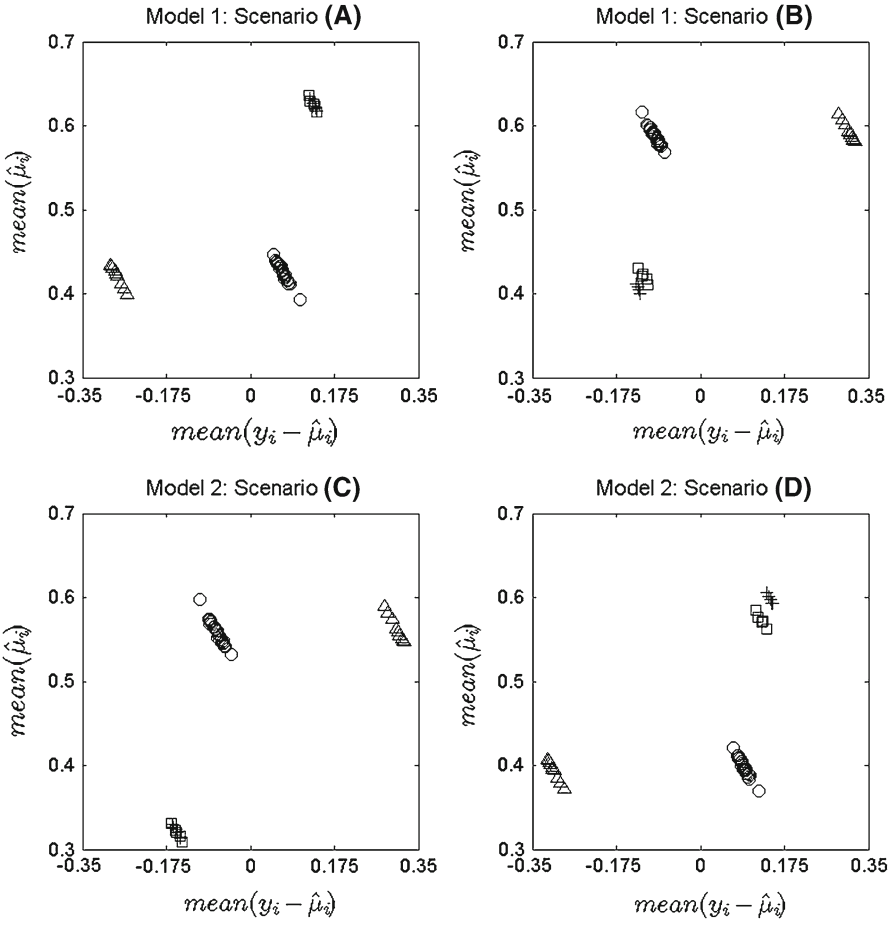

Fig. 1 Scatter plots of the averages of $2000\widehat{\mu}_i$ against $2000(y_i - \widehat{\mu}_i)$ for each outlier scenario for sample size $n = 50$. *Triangle* indexes the outlying observations that have unusual residuals; *square* indexes the outlying data points that are influential on $\widehat{\beta}_2$; *plus* indexes the outlying data points that have an influence on $\widehat{\beta}_3$; *circle* indexes the usual data points

Two thousand simulation runs are carried out for each case, with the $x_{i,2}$'s and $x_{i,3}$'s being regenerated after every 50 simulations. Figure 1 shows scatter plots of the averages of $2000\widehat{\mu}_i$ against $2000(y_i - \widehat{\mu}_i)$ for each outlier scenario for sample size $n = 50$. Because the results of sample sizes $n = 50$ and 150 are similar, we omit the results for $n = 150$ to reduce space. From Fig. 1, it is evident that the $mean(y_i - \widehat{\mu}_i)$ values of the outlying observations that have unusual residuals are greater or less than that of the usual observations and the outlying data points that are influential on $\widehat{\beta}_2$ or $\widehat{\beta}_3$. On the contrary, the $mean(\widehat{\mu}_i)$ values of the outlying data points that have an influence on $\widehat{\beta}_2$ or $\widehat{\beta}_3$ are bigger or lower than that of the usual data points and the outlying observations that have extreme residuals. The indication reveals that the former have extremely unusual residuals but they don't have significant impacts on $\widehat{\mu}_i$, whereas the latter have influences on $\widehat{\mu}_i$ but they don't have extraordinarily unusual residuals.

Tables 1, 2, 3 and 4 display the averages of the proportions of observation $i, i = 1, \ldots, 0.2n$, observation $i, i = 0.2n + 1, \ldots, 0.3n$, and observation $i, i = 0.3n + 1, \ldots, 0.4n$, claimed by the single-case and group deletion diagnostics as outlying observations in the two thousand simulated data sets. Also displayed are the averages of the proportions of observation $i, i = 1, \ldots, n$, identified as outliers in the 2000 replications, in which no observations are generated as outlying observations in the simulated data sets. In addition, Tables 1, 2, 3 and 4 also exhibit the results based on the different diagnostic criteria, in order to compare the results from the suggested cut-off values.

From Tables 1, 2, 3 and 4, it is clear that, in the cases where no observations are generated as outliers in the simulated data sets and the diagnostic criteria are based on the suggested cut-off values, the averages of the proportions of observation $i, i = 1, \ldots, n$, detected by the single-case and group deletion diagnostics as the unusual observations in the 2000 replications gradually approach 0.05 as the sample size n increases. The indications imply that, on the basis of the suggested cut-off values, the probability that the usual observations are misdiagnosed as unusual observations by the single-case and group deletion diagnostics is around 0.05. This, in turn, means that, in the cases where no observations are generated as outliers in the simulated data sets, the results from the single-case and group deletion diagnostics, based on the suggested cut-off values, are reasonable.

On the other hand, from Tables 1, 2, 3 and 4, it is clear that, in the cases with observations generated as outliers in the simulated data sets, the group deletion diagnostics are more effective than the single-case deletion diagnostics, in picking up the outliers from the data sets. As in the case in Model 2 with $n = 150$ under the outlier scenario (C), when the cut-off values of DFBETAS_{i2} and GDFBETAS_{i2} are considered by $2/\sqrt{n}$ and $2c(s)/\sqrt{s}$, respectively, the average of the proportions of observation $i, i = 31, \ldots, 45$, highlighted as outlying data points in the 2000 simulations by GDFBETAS_{i2} is 0.9438, bigger than the corresponding result 0.3661 from DFBETAS_{i2}. Similar results are also obtained by comparing differences between the results from GDFBETAS_{i3} and DFBETAS_{i3} or from GLD_i and LD_i, etc.

From Tables 1, 2, 3 and 4, it is also noted that SWR_i and GSWR_i identify outlying observations that have extra unusual residuals as atypical observations, whereas the outlying data points that have influences on $\widehat{\beta}_2$ or $\widehat{\beta}_3$ are not flagged by SWR_i and GSWR_i as unusual data points. This is because, the latter don't have enough extreme

Table 1 Model 1: logit $(\mu_i) = -0.01 - 0.01x_{i,2} - 0.01x_{i,3}, i = 1, \ldots, 50$

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–50	Observations 1–10	Observations 11–15	Observations 16–20	Observations 21–50
Scenario (A)						
SWR_i	2	0.0518	**0.3205**	0.0116	0.0109	0.0008
$GSWR_i$	$2c(s)$	0.0424	**0.3434**	0.0037	0.0037	0.0002
LD_i	$4p/n$	0.0658	**0.1027**	**0.1601**	**0.1562**	0.0033
GLD_i	$4c^2(s)p/s$	0.0488	0.0431	**0.2786**	**0.2854**	0.0005
$DFFITS_i$	$2\sqrt{p/n}$	0.0714	**0.1319**	**0.1616**	**0.1587**	0.0004
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0623	**0.1033**	**0.4049**	**0.3972**	0.0011
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0810	0.0898	**0.3769**	0.0134	0.0093
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0787	**0.1202**	**0.8435**	0.0061	0.0019
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0816	0.0855	0.0205	**0.3829**	0.0126
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0800	**0.1100**	0.0136	**0.8257**	0.0024
SWR_i	2.5	0.0142	**0.1235**	0.0015	0.0008	0.0000
$GSWR_i$	$2.5c(s)$	0.0131	**0.1228**	0.0012	0.0006	0.0000
LD_i	$6.25p/n$	0.0259	0.0203	0.0728	0.0639	0.0002
GLD_i	$6.25c^2(s)p/s$	0.0225	0.0142	**0.1096**	**0.1054**	0.0001
$DFFITS_i$	$2.5\sqrt{p/n}$	0.0307	0.0337	0.0767	0.0672	0.0006
$GDFFITS_i$	$2.5c(s)\sqrt{p/s}$	0.0288	0.0309	**0.2733**	**0.2791**	0.0001
$DFBETAS_{i2}$	$2.5/\sqrt{n}$	0.0425	0.0344	**0.2570**	0.0044	0.0020
$GDFBETAS_{i2}$	$2.5c(s)/\sqrt{s}$	0.0423	0.0458	**0.7400**	0.0018	0.0003
$DFBETAS_{i3}$	$2.5/\sqrt{n}$	0.0446	0.0319	0.0065	**0.2606**	0.0035
$GDFBETAS_{i3}$	$2.5c(s)/\sqrt{s}$	0.0443	0.0447	0.0039	**0.7312**	0.0009
Scenario (B)						
SWR_i	2	0.0518	**0.3861**	0.0020	0.0016	0.0002
$GSWR_i$	$2c(s)$	0.0424	**0.4485**	0.0004	0.0004	0.0000
LD_i	$4p/n$	0.0658	**0.1426**	0.0742	0.0819	0.0014
GLD_i	$4c^2(s)p/s$	0.0488	**0.1004**	**0.1055**	**0.1483**	0.0005
$DFFITS_i$	$2\sqrt{p/n}$	0.0714	**0.1727**	0.0755	0.0845	0.0017
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0623	**0.1876**	**0.1569**	**0.2074**	0.0004
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0810	**0.1213**	**0.2585**	0.0045	0.0056
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0787	**0.1602**	**0.6769**	0.0019	0.0013
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0816	**0.1131**	0.0069	**0.2886**	0.0081
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0800	**0.1470**	0.0045	**0.7397**	0.0014
SWR_i	2.5	0.0142	**0.1609**	0.0001	0.0000	0.0000
$GSWR_i$	$2.5c(s)$	0.0131	**0.1726**	0.0001	0.0000	0.0000
LD_i	$6.25p/n$	0.0259	0.0351	0.0254	0.0243	0.0002
GLD_i	$6.25c^2(s)p/s$	0.0225	0.0317	0.0322	0.0350	0.0002

Table 1 continued

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–50	Observations 1–10	Observations 11–15	Observations 16–20	Observations 21–50
DFFITS_i	$2.5\sqrt{p/n}$	0.0307	0.0519	0.0266	0.0260	0.0003
GDFFITS_i	$2.5c(s)\sqrt{p/s}$	0.0288	0.0531	0.0744	**0.1054**	0.0002
DFBETAS_{i2}	$2.5/\sqrt{n}$	0.0425	0.0537	**0.1464**	0.0008	0.0009
GDFBETAS_{i2}	$2.5c(s)/\sqrt{s}$	0.0423	0.0683	**0.5066**	0.0003	0.0002
DFBETAS_{i3}	$2.5/\sqrt{n}$	0.0446	0.0461	0.0023	**0.1681**	0.0019
GDFBETAS_{i3}	$2.5c(s)/\sqrt{s}$	0.0443	0.0625	0.0011	**0.5812**	0.0006

The diagnostic results based on single-case and group deletion diagnostics are larger than 0.1, which are respectively highlighted as the significance of bold and underline values

[a] observations 1–10 have unusual residuals; observations 11–15 and 16–20 respectively have influences on $\widehat{\beta}_2$ and $\widehat{\beta}_3$; observations 21–50 are usual data points

residuals of $y_i - \widehat{\mu}_i$, as the results shown in Fig. 1, in which the $mean(y_i - \widehat{\mu}_i)$ values of the latter are closer to that of the usual observations, in comparison with the former. On the other hand, it is seen that due to the latter generated as having impacts on regression parameter estimates, they are more easily identified by DFBETAS_i and GDFBETAS_i as unusual data points.

In addition, it is observed from Tables 1, 2, 3 and 4 that, in general, the averages of the proportions of the outlying data points that have influences on $\widehat{\beta}_2$ or $\widehat{\beta}_3$ detected as unusual data points are higher than that of the outlying observations that have unusual residuals. In other words, it is observed that the latter are identified harder than the former. This is because, the latter are not outlying enough. As compared the results from Model 2 under the outlier scenarios (C) and (D), it is shown that the latter become easier to be identified when their residuals become more excessive.

Obviously, the group deletion diagnostics are able to provide the more valid inferences in the regression diagnostic analyses, under the data sets with multiple outliers.

6 Examples

Two practical applications are presented in this section to illustrate the usefulness of the proposed group deletion diagnostic measures.

6.1 Example 1: Reading accuracy data

The first application uses the data analyzed by Espinheira et al. (2008a) from 44 children (19 children with dyslexia and 25 controls) recruited from primary schools in the Australian Capital Territory. The ages of the children ranged from eight years five months to twelve years three months. The response (y) gives the scores on a test of reading accuracy, and the explanatory variables represent dyslexia versus non-dyslexia

Table 2 Model 1: logit $(\mu_i) = -0.01 - 0.01x_{i,2} - 0.01x_{i,3}, i = 1, \ldots, 150$

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–150	Observations 1–30	Observations 31–45	Observations 46–60	Observations 61–150
Scenario (A)						
SWR_i	2	0.0481	**0.3096**	0.0088	0.0087	0.0006
$GSWR_i$	$2c(s)$	0.0382	**0.3290**	0.0021	0.0018	0.0001
LD_i	$4p/n$	0.0547	**0.1040**	**0.1173**	**0.1198**	0.0013
GLD_i	$4c^2(s)p/s$	0.0377	0.0512	**0.1680**	**0.1618**	0.0002
$DFFITS_i$	$2\sqrt{p/n}$	0.0566	**0.1120**	**0.1183**	**0.1206**	0.0015
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0470	0.0765	**0.3752**	**0.3575**	0.0003
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0697	0.0858	**0.3299**	0.0096	0.0058
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0628	0.0978	**0.8951**	0.0028	0.0003
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0688	0.0805	0.0103	**0.3301**	0.0062
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0622	0.0902	0.0021	**0.8973**	0.0003
SWR_i	2.5	0.0136	**0.1226**	0.0009	0.0006	0.0000
$GSWR_i$	$2.5c(s)$	0.0121	**0.1235**	0.0005	0.0004	0.0000
LD_i	$6.25p/n$	0.0199	0.0228	0.0460	0.0442	0.0001
GLD_i	$6.25c^2(s)p/s$	0.0161	0.0169	0.0500	0.0471	0.0000
$DFFITS_i$	$2.5\sqrt{p/n}$	0.0213	0.0269	0.0470	0.0455	0.0001
$GDFFITS_i$	$2.5c(s)\sqrt{p/s}$	0.0191	0.0215	**0.1666**	**0.1497**	0.0001
$DFBETAS_{i2}$	$2.5/\sqrt{n}$	0.0345	0.0307	**0.2064**	0.0023	0.0008
$GDFBETAS_{i2}$	$2.5c(s)/\sqrt{s}$	0.0320	0.0342	**0.7709**	0.0007	0.0001
$DFBETAS_{i3}$	$2.5/\sqrt{n}$	0.0343	0.0290	0.0025	**0.2067**	0.0008
$GDFBETAS_{i3}$	$2.5c(s)/\sqrt{s}$	0.0319	0.0317	0.0005	**0.7838**	0.0001
Scenario (B)						
SWR_i	2	0.0481	**0.3715**	0.0009	0.0012	0.0001
$GSWR_i$	$2c(s)$	0.0382	**0.4288**	0.0000	0.0001	0.0000
LD_i	$4p/n$	0.0547	**0.1429**	0.0421	0.0503	0.0004
GLD_i	$4c^2(s)p/s$	0.0377	**0.1127**	0.0315	0.0039	0.0001
$DFFITS_i$	$2\sqrt{p/n}$	0.0566	**0.1519**	0.0426	0.0507	0.0005
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0470	**0.1605**	0.0591	**0.1038**	0.0001
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0697	**0.1176**	**0.1987**	0.0031	0.0028
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0628	**0.1383**	**0.6972**	0.0008	0.0002
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0688	**0.1060**	0.0024	**0.2222**	0.0033
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0622	**0.1241**	0.0006	**0.7654**	0.0004
SWR_i	2.5	0.0136	**0.1588**	0.0000	0.0001	0.0000
$GSWR_i$	$2.5c(s)$	0.0121	**0.1724**	0.0000	0.0000	0.0000
LD_i	$6.25p/n$	0.0199	0.0369	0.0100	0.0128	0.0000
GLD_i	$6.25c^2(s)p/s$	0.0161	0.0338	0.0092	0.0114	0.0001

Table 2 continued

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–150	Observations 1–30	Observations 31–45	Observations 46–60	Observations 61–150
DFFITS_i	$2.5\sqrt{p/n}$	0.0213	0.0424	0.0104	0.0131	0.0000
GDFFITS_i	$2.5c(s)\sqrt{p/s}$	0.0191	0.0428	0.0133	0.0247	0.0000
DFBETAS_{i2}	$2.5/\sqrt{n}$	0.0345	0.0487	0.0955	0.0005	0.0002
GDFBETAS_{i2}	$2.5c(s)/\sqrt{s}$	0.0320	0.0541	**0.4334**	0.0004	0.0000
DFBETAS_{i3}	$2.5/\sqrt{n}$	0.0343	0.0430	0.0003	**0.1122**	0.0003
GDFBETAS_{i3}	$2.5c(s)/\sqrt{s}$	0.0319	0.0471	0.0001	**0.5277**	0.0001

The diagnostic results based on single-case and group deletion diagnostics are larger than 0.1, which are respectively highlighted as the significance of bold and underline values

[a] observations 1–30 have unusual residuals; observations 31–45 and 46–60 respectively have influences on $\widehat{\beta}_2$ and $\widehat{\beta}_3$; observations 61–150 are usual data points

status (x_2), non-verbal IQ scores converted to z-scores (x_3) and an interaction variable (x_4). The variable x_2 is coded as 1 when the child is dyslexic and otherwise it is coded as -1. The non-dyslexic readers' mean accuracy score is 0.900 whereas the mean for readers who have dyslexia is 0.606. The overall mean score is 0.773. Following the suggestions in Espinheira et al. (2008a), we analyze the data set using the GBLM with the logit link function as

$$\text{logit}\,(\mu_i) = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}, \quad i = 1, \ldots, 44.$$

Table 5 displays diagnostic results based on the single-case and group deletion diagnostics, respectively. Looking at Table 5, some interesting findings are as follows. First, from GSWR_i and SWR_i, it is observed that GSWR_i suggests observations 8, 9, 15 and 22 as the outlying observations that have unusual residuals, whereas, among them, only observation 8 is highlighted by SWR_i as an unusual observation on the residual. Then comparing the results from LD_i and GLD_i, it is noted that LD_i detects observations 6 and 8 as the outlying data points that have an impact on the regression parameter estimates, $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\beta}_4$, while GLD_i identifies observations 8 and 15 as the atypical observations that are influential on $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\beta}_4$. The joint deletion of observations 6 and 8 shows that the relative changes in $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\beta}_4$ are -1.773, -2.761, 32.77 and 24.36 %, whereas the relative changes in $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\beta}_4$ are -11.32, -16.03, 117.5 and 86.82 % with observations 8 and 15 deleted. The indication implies that observations 8 and 15 jointly have more powerful influences on the regression parameter estimates, in contrast with observations 6 and 8.

Similarly, comparing the results from DFFITS_i and GDFFITS_i, it is also observed that GDFFITS_i detects observations 5, 8, 9, 15 and 22 as the atypical points on the weighted fits. Among these atypical points, observations 5, 9, 15 and 22 are obscured by DFFITS_i. Jointly removing observations 5, 8, 9, 15 and 22 from the data results in the apparently relative variations, -28.10, -39.86, 269.1 and 199.0 % in $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\beta}_4$.

Table 3 Model 2: $\Phi(\mu_i) = -0.01 - 0.01x_{i,2} - 0.01x_{i,3}, i = 1, \ldots, 50$

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–50	Observations 1–10	Observations 11–15	Observations 16–20	Observations 21–50
Scenario (C)						
SWR_i	2	0.0517	**0.3082**	0.0193	0.0182	0.0003
$GSWR_i$	$2c(s)$	0.0423	**0.3242**	0.0087	0.0082	0.0001
LD_i	$4p/n$	0.0657	0.0828	**0.1917**	**0.1907**	0.0030
GLD_i	$4c^2(s)p/s$	0.0489	0.0231	**0.3460**	**0.3614**	0.0003
$DFFITS_i$	$2\sqrt{p/n}$	0.0714	**0.1488**	**0.1099**	**0.1040**	0.0037
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0624	**0.1440**	**0.2729**	**0.2739**	0.0010
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0425	0.0761	**0.4132**	0.0189	0.0091
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0424	**0.1017**	**0.8840**	0.0079	0.0015
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0815	0.0723	0.0295	**0.4245**	0.0122
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0799	0.0981	0.0173	**0.8841**	0.0016
SWR_i	2.5	0.0142	**0.1084**	0.0028	0.0017	0.0000
$GSWR_i$	$2.5c(s)$	0.0132	**0.1051**	0.0025	0.0014	0.0000
LD_i	$6.25p/n$	0.0258	0.0136	0.0908	0.0856	0.0005
GLD_i	$6.25c^2(s)p/s$	0.0225	0.0080	**0.1467**	**0.1434**	0.0002
$DFFITS_i$	$2.5\sqrt{p/n}$	0.0307	0.0403	0.0451	0.0384	0.0006
$GDFFITS_i$	$2.5c(s)\sqrt{p/s}$	0.0287	0.0393	**0.1636**	**0.1530**	0.0003
$DFBETAS_{i2}$	$2.5/\sqrt{n}$	0.0811	0.0260	**0.2940**	0.0064	0.0019
$GDFBETAS_{i2}$	$2.5c(s)/\sqrt{s}$	0.0787	0.0358	**0.8121**	0.0013	0.0002
$DFBETAS_{i3}$	$2.5/\sqrt{n}$	0.0446	0.0248	0.0109	**0.3007**	0.0035
$GDFBETAS_{i3}$	$2.5c(s)/\sqrt{s}$	0.0442	0.0369	0.0049	**0.8163**	0.0005
Scenario (D)						
SWR_i	2	0.0517	**0.3891**	0.0024	0.0025	0.0001
$GSWR_i$	$2c(s)$	0.0423	**0.4513**	0.0002	0.0002	0.0000
LD_i	$4p/n$	0.0657	**0.1448**	0.0843	**0.1009**	0.0018
GLD_i	$4c^2(s)p/s$	0.0489	0.0923	**0.1229**	**0.2048**	0.0004
$DFFITS_i$	$2\sqrt{p/n}$	0.0714	**0.1933**	0.0553	0.0601	0.0020
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0624	**0.2295**	0.0746	**0.1109**	0.0005
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0425	**0.1192**	**0.2803**	0.0071	0.0056
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0424	**0.1537**	**0.7037**	0.0036	0.0014
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0815	**0.1069**	0.0083	**0.3224**	0.0087
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0799	**0.1431**	0.0045	**0.7903**	0.0013
SWR_i	2.5	0.0142	**0.1704**	0.0001	0.0001	0.0000
$GSWR_i$	$2.5c(s)$	0.0132	**0.1820**	0.0001	0.0000	0.0000
LD_i	$6.25p/n$	0.0258	0.0368	0.0290	0.0314	0.0001
GLD_i	$6.25c^2(s)p/s$	0.0225	0.0317	0.0386	0.0535	0.0001

Table 3 continued

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–50	Observations 1–10	Observations 11–15	Observations 16–20	Observations 21–50
DFFITS_i	$2.5\sqrt{p/n}$	0.0307	0.0619	0.0168	0.0155	0.0001
GDFFITS_i	$2.5c(s)\sqrt{p/s}$	0.0287	0.0661	0.0323	0.0457	0.0001
DFBETAS_{i2}	$2.5/\sqrt{n}$	0.0811	0.0533	**0.1638**	0.0015	0.0008
GDFBETAS_{i2}	$2.5c(s)/\sqrt{s}$	0.0787	0.0663	**0.5440**	0.0005	0.0001
DFBETAS_{i3}	$2.5/\sqrt{n}$	0.0446	0.0425	0.0027	**0.1996**	0.0023
GDFBETAS_{i3}	$2.5c(s)/\sqrt{s}$	0.0442	0.0578	0.0010	**0.6485**	0.0005

The diagnostic results based on single-case and group deletion diagnostics are larger than 0.1, which are respectively highlighted as the significance of bold and underline values

[a] observations 1–10 have unusual residuals; observations 11–15 and 16–20 respectively have influences on $\widehat{\beta}_2$ and $\widehat{\beta}_3$; observations 21–50 are usual data points

On the other hand, DFBETAS_{i3} and DFBETAS_{i4} detect observations 6, 8 and 15 as having a large impact on $\widehat{\beta}_3$ and $\widehat{\beta}_4$, while GDFBETAS_{i3} and GDFBETAS_{i4} spot observations 5, 8, 9, 15 and 22 as influential on the results for $\widehat{\beta}_3$ and $\widehat{\beta}_4$. The exclusion of observations 5, 8, 9, 15 and 22 causes an apparent change 269.1 and 199.0 % in $\widehat{\beta}_3$ and $\widehat{\beta}_4$, whereas the relative change in $\widehat{\beta}_3$ and $\widehat{\beta}_4$ are 86.31 and 63.88 % with observations 6, 8 and 15 omitted together. This indicates that observations 5, 8, 9, 15 and 22 have a relatively larger effect on $\widehat{\beta}_3$ and $\widehat{\beta}_4$, compared with observations 6, 8 and 15. In addition, it should be explained that, in this example, we don't consider the results of GDFBETAS_{i2} and DFBETAS_{i2}. This is because, x_2 is a binary variable whose value equals -1 or 1. Thus we are not very interested that which observation is influential on $\widehat{\beta}_2$.

Evidently, in reading accuracy data, the group deletion diagnostics detect some multiple outlying observations that are influential on the results of the regression analysis, while these unusual observations are ignored by the single-case deletion diagnostics.

6.2 Example 2: Stress, depression, and anxiety

The second example uses data collected from a sample of 166 nonclinical women in Townsville, Queensland, Australia. The data were analyzed by Smithson and Verkuilen (2006) using beta regression. The response variable (y) represents the scores on a test of anxiety symptoms, and the explanatory variable (x_2) gives the corresponding scores on the test of stress symptoms. We follow Smithson and Verkuilen (2006) in considering the model

$$\text{logit}\,(\mu_i) = \beta_1 + \beta_2 x_{i,2}, \quad i = 1, \ldots, 166.$$

Table 6 presents diagnostic results based on the single-case and group deletion diagnostics, respectively. Careful inspection of Table 6 indicates some interesting findings as follows.

Table 4 Model 2: $\Phi(\mu_i) = -0.01 - 0.01x_{i,2} - 0.01x_{i,3}, i = 1, \ldots, 150$

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–150	Observations 1–30	Observations 31–45	Observations 46–60	Observations 61–150
Scenario (C)						
SWR_i	2	0.0481	**0.2960**	0.0147	0.0146	0.0002
$GSWR_i$	$2c(s)$	0.0381	**0.3111**	0.0038	0.0049	0.0000
LD_i	$4p/n$	0.0547	0.0843	**0.1448**	**0.1459**	0.0011
GLD_i	$4c^2(s)p/s$	0.0377	0.0277	**0.2434**	**0.2172**	0.0000
$DFFITS_i$	$2\sqrt{p/n}$	0.0566	**0.1267**	0.0716	0.0723	0.0012
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0470	**0.1153**	**0.1806**	**0.1648**	0.0004
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0700	0.0731	**0.3661**	0.0160	0.0050
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0629	0.0807	**0.9438**	0.0023	0.0001
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0689	0.0677	0.0159	**0.3636**	0.0054
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0623	0.0744	0.0027	**0.9466**	0.0001
SWR_i	2.5	0.0135	**0.1088**	0.0019	0.0022	0.0000
$GSWR_i$	$2.5c(s)$	0.0120	**0.1085**	0.0011	0.0010	0.0000
LD_i	$6.25p/n$	0.0200	0.0154	0.0606	0.0608	0.0000
GLD_i	$6.25c^2(s)p/s$	0.0161	0.0098	0.0717	0.0675	0.0000
$DFFITS_i$	$2.5\sqrt{p/n}$	0.0213	0.0310	0.0222	0.0212	0.0001
$GDFFITS_i$	$2.5c(s)\sqrt{p/s}$	0.0191	0.0297	0.0458	0.0473	0.0000
$DFBETAS_{i2}$	$2.5/\sqrt{n}$	0.0345	0.0230	**0.2409**	0.0041	0.0005
$GDFBETAS_{i2}$	$2.5c(s)/\sqrt{s}$	0.0321	0.0252	**0.8718**	0.0008	0.0001
$DFBETAS_{i3}$	$2.5/\sqrt{n}$	0.0343	0.0220	0.0043	**0.2384**	0.0005
$GDFBETAS_{i3}$	$2.5c(s)/\sqrt{s}$	0.0319	0.0231	0.0005	**0.8763**	0.0000
Scenario (D)						
SWR_i	2	0.0481	**0.3723**	0.0010	0.0022	0.0001
$GSWR_i$	$2c(s)$	0.0381	**0.4271**	0.0000	0.0002	0.0000
LD_i	$4p/n$	0.0547	**0.1450**	0.0487	0.0657	0.0006
GLD_i	$4c^2(s)p/s$	0.0377	**0.1092**	0.0355	0.0650	0.0001
$DFFITS_i$	$2\sqrt{p/n}$	0.0566	**0.1698**	0.0259	0.0316	0.0007
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	0.0470	**0.1940**	0.0208	0.0353	0.0002
$DFBETAS_{i2}$	$2/\sqrt{n}$	0.0700	**0.1156**	**0.2196**	0.0044	0.0029
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	0.0629	**0.1370**	**0.7267**	0.0001	0.0002
$DFBETAS_{i3}$	$2/\sqrt{n}$	0.0689	**0.1013**	0.0030	**0.2653**	0.0036
$GDFBETAS_{i3}$	$2c(s)/\sqrt{s}$	0.0623	**0.1196**	0.0006	**0.8432**	0.0002
SWR_i	2.5	0.0135	**0.1678**	0.0000	0.0001	0.0000
$GSWR_i$	$2.5c(s)$	0.0120	**0.1826**	0.0000	0.0001	0.0000
LD_i	$6.25p/n$	0.0200	0.0398	0.0118	0.0167	0.0000
GLD_i	$6.25c^2(s)p/s$	0.0161	0.0355	0.0110	0.0154	0.0000

Table 4 continued

Diagnostic	Cut-off value	Data without outlying observations	Data with outlying observations[a]			
		Observations 1–150	Observations 1–30	Observations 31–45	Observations 46–60	Observations 61–150
DFFITS_i	$2.5\sqrt{p/n}$	0.0213	0.0514	0.0050	0.0006	0.0000
GDFFITS_i	$2.5c(s)\sqrt{p/s}$	0.0191	0.0539	0.0044	0.0078	0.0000
DFBETAS_{i2}	$2.5/\sqrt{n}$	0.0345	0.0479	**0.1091**	0.0008	0.0003
GDFBETAS_{i2}	$2.5c(s)/\sqrt{s}$	0.0321	0.0530	**0.4839**	0.0004	0.0000
DFBETAS_{i3}	$2.5/\sqrt{n}$	0.0343	0.0408	0.0004	**0.1401**	0.0003
GDFBETAS_{i3}	$2.5c(s)/\sqrt{s}$	0.0319	0.0447	0.0002	**0.6547**	0.0000

The diagnostic results based on single-case and group deletion diagnostics are larger than 0.1, which are respectively highlighted as the significance of bold and underline values

[a] Observations 1–30 have unusual residuals; observations 31–45 and 46–60 respectively have influences on $\widehat{\beta}_2$ and $\widehat{\beta}_3$; observations 61–150 are usual data points

Table 5 The single-case and group deletion diagnostics for example 1

Diagnostic	Cut-off value	Case i claimed as an outlier	Cut-off value	Case i claimed as an outlier	Cut-off value	Case i claimed as an outlier
SWR_i	2	8	2.5	No observations[a]	3	No observations[a]
GSWR_i	$2c(s)$	8, 9, 15, 22	$2.5c(s)$	8	$3c(s)$	No observations[a]
LD_i	$4p/n$	6, 8	$6.25p/n$	8	$9p/n$	8
GLD_i	$4c^2(s)p/s$	8, 15	$6.25c^2(s)p/s$	8	$9c^2(s)p/s$	8
DFFITS_i	$2\sqrt{p/n}$	6, 8	$2.5\sqrt{p/n}$	8	$3\sqrt{p/n}$	8
GDFFITS_i	$2c(s)\sqrt{p/s}$	5, 8, 9, 15, 22	$2.5c(s)\sqrt{p/s}$	8, 15, 22	$3c(s)\sqrt{p/s}$	8
DFBETAS_{i3}	$2/\sqrt{n}$	6, 8, 15	$2.5/\sqrt{n}$	6, 8	$3/\sqrt{n}$	8
GDFBETAS_{i3}	$2c(s)/\sqrt{s}$	5, 8, 9, 15, 22	$2.5c(s)/\sqrt{s}$	5, 8, 9, 15, 22	$3c(s)/\sqrt{s}$	8, 9, 15, 22
DFBETAS_{i4}	$2/\sqrt{n}$	6, 8, 15	$2.5/\sqrt{n}$	6, 8	$3/\sqrt{n}$	8
GDFBETAS_{i4}	$2c(s)/\sqrt{s}$	5, 8, 9, 15, 22	$2.5c(s)/\sqrt{s}$	5, 8, 9, 15, 22	$3c(s)/\sqrt{s}$	8, 9, 15, 22

[a] No observations are claimed as outliers

From the results of SWR_i and GSWR_i, it is shown that diagnostic results based on SWR_i and GSWR_i are similar. Observations 10, 89, 116 and 136 are flagged by SWR_i and GSWR_i as the outlying observations with extra large or small residuals.

Similarly, from the results of LD_i and GLD_i, it is shown that GLD_i and LD_i suggest observations 10, 55, 77, 89, 116, 125, 132, 151, 152 and 164 as the outlying data points that are influential on the regression parameter estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$. It is also interesting to note that, on the basis of cut-off values $9c^2(s)p/s$ and $9p/n$, GLD_i suggests cases 89, 116, 151 and 152 as having an influence on $\widehat{\beta}_1$ and $\widehat{\beta}_2$, while LD_i identifies cases 55, 89, 116 and 152 as influential on the results for $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The joint deletion of observations 89, 116, 151 and 152 causes the relative changes in $\widehat{\beta}_1$ and $\widehat{\beta}_2$ as 5.21 and 13.80 %, whereas the relative changes in $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are 3.47 and 7.74 % with

Table 6 The single-case and group deletion diagnostics for example 2

Diagnostic	Cut-off value	Case i claimed as an outlier	Cut-off value	Case i claimed as an outlier	Cut-off value	Case i claimed as an outlier
SWR_i	2	10, 89, 116, 136	2.5	10, 89, 116	3	89
$GSWR_i$	$2c(s)$	10, 89, 116, 136	$2.5c(s)$	10, 89, 116	$3c(s)$	89
LD_i	$4p/n$	10, 55, 77, 89, 116, 125, 132, 151, 152, 164	$6.25p/n$	55, 77, 89, 116, 125, 132, 151, 152, 164	$9p/n$	55, 89, 116, 152
GLD_i	$4c^2(s)p/s$	10, 55, 77, 89, 116, 125, 132, 151, 152, 164	$6.25c^2(s)p/s$	55, 77, 89, 116, 125, 132, 152, 164	$9c^2(s)p/s$	89, 116, 151, 152
$DFFITS_i$	$2\sqrt{p/n}$	10, 55, 77, 89, 116, 125, 132, 151, 152, 164	$2.5\sqrt{p/n}$	55, 77, 89, 116, 125, 132, 151, 152, 164	$3\sqrt{p/n}$	55, 89, 116, 152
$GDFFITS_i$	$2c(s)\sqrt{p/s}$	10, 51, 55, 77, 89, 116, 125, 132, 133, 151, 152, 164	$2.5c(s)\sqrt{p/s}$	10, 55, 77, 89, 116, 125, 132, 151, 152, 164	$3c(s)\sqrt{p/s}$	55, 77, 89, 116, 125, 132, 152, 164
$DFBETAS_{i2}$	$2/\sqrt{n}$	55, 77, 89, 116, 125, 132, 151, 152, 164	$2.5/\sqrt{n}$	55, 77, 89, 116, 125, 132, 151, 152, 164	$3/\sqrt{n}$	55, 77, 89, 116, 125, 132, 151, 152, 164
$GDFBETAS_{i2}$	$2c(s)/\sqrt{s}$	10, 51, 55, 77, 89, 116, 117, 125, 132, 133, 151, 152, 164	$2.5c(s)/\sqrt{s}$	51, 55, 77, 89, 116, 125, 132, 151, 152, 164	$3c(s)/\sqrt{s}$	55, 77, 89, 116, 125, 132, 151, 152, 164

observations 55, 89, 116 and 152 eliminated. The indication shows that GLD_i is more effective than LD_i, in pointing out the outlying data points that have jointly bigger impacts on $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

From the results of $DFFITS_i$ and $GDFFITS_i$, it is observed that $DFFITS_i$ misses observations 51 and 133 that are discovered by $GDFFITS_i$ as the atypical points on the weighted fits. The joint deletion of observations 10, 55, 77, 89, 116, 125, 132, 151, 152 and 164 leads to the relative variations in $\widehat{\beta}_1$ and $\widehat{\beta}_2$ as 2.60 and 2.00 %, whereas the joint deletion of observations 10, 51, 55, 77, 89, 116, 125, 132, 133, 151, 152 and 164 leads to the relative variations in $\widehat{\beta}_1$ and $\widehat{\beta}_2$ as 2.76 and −1.93 %. It is evident that observations 51 and 133 that are omitted by $DFFITS_i$ are influential, because deleting the atypical points with observations 51 and 133 results in a relative change of $\widehat{\beta}_2$ from 2.00 (positive) to −1.93 % (negative).

From the results of $DFBETAS_{i2}$ and $GDFBETAS_{i2}$, it is shown that $GDFBETAS_{i2}$ detects observations 10, 51, 55, 77, 89, 116, 117, 125, 132, 133, 151, 152 and 164 as having a large impact on $\widehat{\beta}_2$, while $DFBETAS_{i2}$ suggests observations 55, 77, 89, 116, 125, 132, 151, 152 and 164 as influential on the results for $\widehat{\beta}_2$. When observations 10, 51, 55, 77, 89, 116, 117, 125, 132, 133, 151, 152 and 164 are simultaneously discarded, $\widehat{\beta}_2$ has a positive jump of 0.83 %, whereas the relative change in $\widehat{\beta}_2$ is a 1.51 % reduction, when observations 55, 77, 89, 116, 125, 132, 151, 152 and 164 are

jointly deleted. This indicates that the outlying data points, observations 10, 51, 117 and 133, are influential on $\widehat{\beta}_2$, but they are not found by DFBETAS$_{i2}$.

Clearly, in this illustration, the group deletion diagnostics suggest some multiple outlying observations that have joint effects on the regression outcome for further considerations, while these potential outlying data points are camouflaged by the single-case deletion diagnostics.

7 Conclusions

In this article, we provide a diagnostic way for the identification of multiple outliers in GBLMs. We suggest the group deletion diagnostic measures, GSWR$_i$, GLD$_i$, GDFFITS$_i$ and GDFBETAS$_i$, respectively, and propose a test procedure for detecting multiple outlying observations using these. Simulation studies and analysis of two practical examples show that our proposed methods can assist the analyst in detecting multiple outlying observations in GBLMs.

Finally, we note that the GBLMs with the constant precision parameter were recently improved by Simas et al. (2010) who allowed a regression structure for the precision parameter. In future work we will extend the group deletion diagnostic techniques to beta regression models with regressors for the precision parameter.

Acknowledgments The author is deeply indebted to the associate editor and two referees for their helpful comments and suggestions that substantially improve this present version of the paper.

References

Atkinson AC (1994) Fast very robust methods for the detection of multiple outliers. J Am Stat Assoc 89:1329–1339

Atkinson AC, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Atkinson AC, Riani M, Cerioli A (2010) The forward search: theory and data analysis (with discussion). J Korean Stat Soc 39:117–134

Belsley DA, Kuh E, Welsch RE (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Cribari-Neto F, Zeileis A (2010) Beta regression in R. J Stat Softw 34:1–24

Espinheira PL, Ferrari SLP, Cribari-Neto F (2008a) On beta regression residuals. J Appl Stat 35:407–419

Espinheira PL, Ferrari SLP, Cribari-Neto F (2008b) Influence diagnostics in beta regression. Comput Stat Data Anal 52:4417–4431

Ferrari SLP, Cribari-Neto F (2004) Beta regression for modeling rates and proportions. J Appl Stat 31: 799–815

Hadi AS, Simonoff JS (1993) Procedures for the identification of multiple outliers in linear models. J Am Stat Assoc 88:1264–1272

Imon AHMR (2005) Identifying multiple influential observations in linear regression. J Appl Stat 32: 929–946

Imon AHMR, Hadi AS (2008) Identification of multiple outliers in logistic regression. Commun Stat Theory Methods 37:1697–1709

Kieschnick R, McCullough BD (2003) Regression analysis of variates observed on (0, 1): percentage, proportions, fractions. Stat Model 3:193–213

Nurunnabi AAM, Imon AHMR, Nasser M (2010) Identification of multiple influential observations in logistic regression. J Appl Stat 37:1605–1624

Simas AB, Barreto-Souza W, Rocha AV (2010) Improved estimators for a general class of beta regression models. Comput Stat Data Anal 54:348–366

Smithson M, Verkuilen J (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol Methods 11:54–71