Journal of Statistics: Advances in Theory and Applications
Volume 8, Number 1, 2012, Pages 37-56

LOCAL INFLUENCE IN COMPOUND-POISSON MODELS: PERTURBING THE MEAN-VARIANCE RELATION

LILA RICCI and PATRICIA ALEGRE

Facultad de Ciencias Exactas
Universidad Nacional de Mar del Plata
Funes 3350, 7600 Mar del Plata
Argentina
e-mail: lricci@mdp.edu.ar

Facultad de Ciencias Económicas
Universidad Nacional de Mar del Plata
Mar del Plata
Argentina

Abstract

Local influence is a useful tool to detect abnormalities in regression models, Cook proposed this method in 1986 for classical regression models and, since then, numerous extensions have been developed. The aim of this paper is to derive methods to asses local influence under various perturbation schemes, for compound-Poisson regression models. These models can be applied to continuous data with positive probability in zero, and they are characterized by the variance function that defines the mean-variance relationship. Formulas are obtained to apply local influence methods for different perturbations and it is of particular interest the perturbation of the parameter that defines the mean-variance relation. These schemes are applied to perturbed data generated by simulations and the sensibility of the method is compared for

2010 Mathematics Subject Classification: 62J20.

Keywords and phrases: compound-Poisson models, local influence.

Received August 18, 2012

different values of the parameters. Finally, a real data set about home expenditures is analyzed and local influence graphics are obtained to detect influential points.

1. Introduction

The main goal of this paper is to adapt local influence methods proposed by Cook [4] for classical linear models and extended by Thomas and Cook [24] for generalized linear models, to compound-Poisson models. The usual perturbation schemes (case, covariables, and response perturbations) are considered and also a new scheme that perturbs the mean-variance relationship is proposed and analyzed.

Local influence method has become an important tool, largely applied in regression analysis. Its goal is to study how sensitive the results of the analysis are to minor perturbations applied on the data or in the model itself. Many applications of the local influence method can be found in the statistical literature for various regression models under different perturbation schemes. Under normal errors, for instance, Lawrence [14] investigated the case of transformed response, Beckman et al. [2] of linear mixed-effect analysis of variance, Tsai and Wu [25] studied the case of auto regressive models, and Molenberghs et al. [15] applied local influence to assess the sensitivity of the dropout process in longitudinal studies. An important extension to generalized linear models was proposed by Thomas and Cook [24] making it possible to consider a wider scenario. Since then numerous articles have considered other contexts, some of them are: restricted generalized linear models (Paula [19]), generalized log-gamma regression models (Ortega et al. [17]), negative binomial (Svetliza and Paula [22]), elliptical t-distributions (Galea et al. [9]), elliptical linear models with longitudinal structure (Osorio et al. [18]), reproductive dispersion models (Tang et al. [23] and Fu et al. [8]), and Poisson inverse Gaussian regression models (Xie and Wei [27]).

In this paper, we consider the problem of assessing local influence in compound-Poisson regression models. These distributions are a subset of Tweedie models (see Tweedie [26] and Jørgensen [12]) characterized by

their mean-variance relationship given by $V(\mu) = \mu^p$, $p \in \mathbb{R} - (0, 1)$ being μ the expected value; given a set of observations, the optimal value for p can be calculated via profile likelihood (Dunn and Smyth [6]) and in this way, one can choose between infinite options for the model. A drawback is that their density functions can not be written in closed form, however, they have simple moment generating functions, so the densities can be evaluated numerically by Fourier inversion of the characteristic functions (Dunn and Smyth [7]). Other remarkable aspects about these models are that, they are exponential dispersion families invariant under change of scale and they are also limit distributions for some exponential dispersion models (Jørgensen [13]). They have been applied, for example, to insurance claims by Smyth and Jørgensen [21], fisheries (Shono [20] and Candy [3]), rainfall prediction (Dunn [5]), and home expenditures (Alegre et al. [1]).

In Section 2, we present the compound-Poisson distributions and derive their first moments as well as the score function and the information matrix. Section 3 is devoted to describe local influence analysis applied to these models, under various perturbation schemes, one of them perturbs the mean-variance relation. Section 4 presents a simulation study and applications to real data. Finally, in Section 5, we elaborate some conclusions.

2. Compound-Poisson Models

Let be independent random variables, N is a Poisson distributed variable, and X_i are have distributions in the exponential dispersion family (Jørgensen [11] and Dunn and Smyth [7]). Let $\mathcal{Y}$ be a random variable defined as follows:

$$\mathcal{Y} = \begin{cases} \sum_{i=1}^{N} X_i, & \text{if } N \neq 0, \\ 0, & \text{if } N = 0, \end{cases}$$

it can be proved that the density function of $\mathcal{Y}$, for $1 < p < 2$ is given by

$$p_p(y, \theta, \phi) = c_p(y, \phi) \exp\left(\frac{\theta y - \kappa_p(\theta)}{\phi}\right); \quad y > 0,$$

$$P(Y = 0) = \exp\left(-\frac{\kappa_p(\theta)}{\phi}\right), \tag{1}$$

where

$$\kappa_p(\theta) = \frac{1}{2-p}((1-p)\theta)^{\frac{p-2}{p-1}}, \tag{2}$$

is the cumulant function and

$$c_p(y, \phi) = \begin{cases} \frac{1}{y}\sum_{k=1}^{\infty} \frac{\kappa_p^k(-\frac{\phi}{y})}{\phi^k \Gamma(-\frac{p-2}{p-1}k)k!}, & \text{when } y > 0, \\ 1, & \text{when } y = 0, \end{cases}$$

being $\theta \in \mathbb{R}^-$ the position parameter and $\phi > 0$ the dispersion parameter. Limit cases are Poisson $(p = 1)$ and gamma $(p = 2)$ models.

The mean and variance of $\mathcal{Y}$ are

$$E(\mathcal{Y}) = \mu = \dot{\kappa}_p(\theta) = ((1-p)\theta)^{\frac{1}{1-p}}, \tag{3}$$

$$\text{Var}(\mathcal{Y}) = \phi\ddot{\kappa}_p(\theta) = \phi\mu^p = \phi V(\mu),$$

where $V(\mu) = \mu^p$ is known as variance function and it characterizes the distribution of $\mathcal{Y}$.

Given a vector of observed values $\mathbf{y} = [y_1 y_2 \ldots y_n]^t$, the log-likelihood is given by

$$L(\mathbf{y}, \boldsymbol{\theta}, \phi) = \sum_{i=1}^{n}\left(\log c_p(y_i, \phi) + \frac{1}{\phi}(y_i\theta_i - \kappa_p(\theta_i))\right). \tag{4}$$

For simplicity, it will be assumed that ϕ is fixed or that it has been estimated externally. Let us consider now a generalized linear regression model (Nelder and Wedderburn [16]) given by

$$g(\boldsymbol{\mu}) = X\boldsymbol{\beta} = \boldsymbol{\theta},$$

where X is the matrix of explanatory variables with dimension $n \times (q+1)$, $\boldsymbol{\beta}$ is the vector of coefficients, and g is a growing and smooth function given by $\dot{\kappa}_p^{-1}(\theta)$, it's named canonical link and it ensures sufficient estimators for $\boldsymbol{\beta}$. Another link function have been proposed by Dunn and Smyth [7], who modifies $\kappa(\theta)$ in such a way that $\mu = 1$ when $\theta = 0$ and by Hardin and Hilbe [10] that define μ^{1-p} as the link function.

The log-likelihood can be re-written in terms of $\boldsymbol{\beta}$ as

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\log c_p(y_i, \phi) + \frac{1}{\phi}\left(y_i \sum_{j=1}^{q+1} x_{ij}\beta_j - \kappa_p\left(\sum_{j=1}^{q+1} x_{ij}\beta_j\right)\right)\right) = \sum_{i=1}^{n} L_i. \quad (5)$$

The components of the score function for $\boldsymbol{\beta}$ are

$$\begin{aligned}\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} &= \frac{1}{\phi}\left(y_i - \frac{\partial \kappa_p(\theta_i)}{\partial \theta_i}\right) x_{ij} \\ &= \frac{1}{\phi}\left(y_i - ((1-p)\theta_i)^{\frac{1}{1-p}}\right) x_{ij} \\ &= \frac{1}{\phi}(y_i - \mu_i)\, x_{ij},\end{aligned}$$

and in matrix form,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \dot{L} = \frac{1}{\phi} X^t(\mathbf{y} - \boldsymbol{\mu}).$$

The observed information matrix for compound-Poisson models with canonical link is given element wise by

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k}\frac{1}{\phi}\sum_{i=1}^{n}\left(y_i - \theta_i^{p/(1-p)}\right) x_{ij}$$

$$= -\frac{1}{\phi}\sum_{i=1}^{n}\frac{\partial\theta_i^{p/(1-p)}}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta_k}x_{ij}$$

$$= -\frac{1}{\phi}\sum_{i=1}^{n}((1-p)\theta_i)^{\frac{p}{1-p}}x_{ji}x_{ki}$$

$$= -\frac{1}{\phi}\sum_{i=1}^{n}\mu_i^p x_{ji}x_{ki},$$

and in matrix form,

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t} = \ddot{L} = -\frac{1}{\phi}X^t\mathrm{diag}(\boldsymbol{\mu}^p)X, \tag{6}$$

where $\mathrm{diag}(\boldsymbol{\mu}^p)$ is a diagonal matrix with μ_i^p in place (i, i).

3. Local Influence

Let us suppose now that data is affected by a perturbation scheme represented by a vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^t$ that belongs to some open region $\Omega \subset \mathbb{R}^n$. Let $L(\boldsymbol{\beta}^{\omega})$ be the log-likelihood of the perturbed model and $\hat{\boldsymbol{\beta}}^{\omega}$ be the corresponding maximum likelihood estimator. Cook [4] proposed to evaluate the influence of $\boldsymbol{\omega}$ on the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, by considering the likelihood displacement given by the function

$$F(\boldsymbol{\omega}) = 2\left(L\left(\hat{\boldsymbol{\beta}}\right) - L\left(\hat{\boldsymbol{\beta}}^{\omega}\right)\right). \tag{7}$$

Using expression (7), the estimators $\hat{\boldsymbol{\beta}}$ obtained for the original model can be compared with those obtained for the perturbed model $\hat{\boldsymbol{\beta}}^{\omega}$. On the other hand, for each scheme, there will be a vector ω_0 in Ω representing no perturbation, in such a way that $L\left(\hat{\boldsymbol{\beta}}^{\omega_0}\right) = L\left(\hat{\boldsymbol{\beta}}\right)$ and $F(\boldsymbol{\omega}_0) = 0$.

The surface defined by $F(\boldsymbol{\omega})$ reaches a minimum at $\boldsymbol{\omega}_0$, for this reason, Cook [4] proposed an influence diagnostic procedure that consists in choosing at that point, those directions $\mathbf{d}$ of greater variation. A first approximation to the surface is given by its tangent plane, but in ω_0, the function has a minimum so the plane is horizontal and gives no information at all. It becomes necessary to work with second order approximations, given by normal curvatures $C(\mathbf{d})$, and then to analyze the unit direction $\mathbf{d}_{\max}$ such that the normal curvature $C_{\max}$ defined by this direction is the one of greater variation. Cook [4] proved that, for linear models, $C(\mathbf{d})$ takes the form

$$C(\mathbf{d}) = \left| \mathbf{d}^t \left. \frac{\partial^2 F(\boldsymbol{\omega})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^t} \right|_{\boldsymbol{\omega}_0} \mathbf{d} \right| = 2 \left| \mathbf{d}^t \Delta^t \left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \right)^{-1} \Delta \mathbf{d} \right|_{\boldsymbol{\omega}_0, \hat{\beta}},$$

where Δ is a $(q+1) \times n$ matrix given by

$$\Delta_{ji} = \frac{1}{\hat{\phi}} \left. \frac{\partial^2 L(\beta \mid \boldsymbol{\omega})}{\partial \beta_j \partial \omega_i} \right|_{\boldsymbol{\omega}_{0i}, \hat{\beta}_j}, \quad 1 \le i \le n, \quad 1 \le j \le q+1, \tag{8}$$

this expression will depend on the perturbation scheme.

Let $C_{\max}$ be the greater eigenvalue of $\Delta^t \hat{\hat{L}}^{-1} \Delta$ and $\mathbf{d}_{\max}$ be the corresponding eigenvector. A plot of the elements of $\mathbf{d}_{\max}$ will reveal those observations that exert a great influence on $L(\boldsymbol{\beta})$ under small perturbations. Generally speaking, to detect influential points, the following steps have to be performed:

(1) Specify a perturbation scheme.

(2) Obtain Δ for the selected perturbation scheme.

(3) Calculate the unit direction of maximum normal curvature $\mathbf{d}_{\max}$.

(4) Make an index plot of $\mathbf{d}_{\max}$.

In the following subsections, some schemes of perturbation to compound-Poisson models will be detailed. First, the usual ones (cases, one covariable and vector of responses) and finally the perturbation of the parameter p, that defines the mean-variance relation, will be analyzed.

3.1. Perturbing cases

Step 1. The elements of $\boldsymbol{\omega}$ can be viewed as weights for each case that perturb the terms of the log-likelihood. The likelihood of the perturbed model is, regarding (4)

$$L(\boldsymbol{\beta}\,|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \log c_p(y_i,\, \phi) + \frac{1}{\phi} \sum_{i=1}^{n} \omega_i (y_i \theta_i^{\omega_i} - \kappa(\theta_i^{\omega_i})). \tag{9}$$

Recall that the first term in (9) will vanish when the expression is differentiated with respect to $\boldsymbol{\beta}$ and the series will then disappear. The no perturbation vector is $\boldsymbol{\omega}_0 = (1, \ldots, 1)^t$.

Step 2. For this kind of perturbation and taking into account (2) and (4), Δ is determined by

$$\begin{aligned} \Delta_{ji} &= \left.\frac{\partial^2 L_i}{\partial \omega_i \partial \beta_j}\right|_{\omega_0, \hat{\beta}} \\ &= \left.\frac{\partial}{\partial \omega_i}\left(\frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}\right)\right|_{\boldsymbol{\omega}_0, \hat{\boldsymbol{\beta}}} \\ &= \left.\frac{1}{\phi}\left(y_i - ((1-p)\theta_i)^{\frac{1}{1-p}}\right) x_{ij}\right|_{\boldsymbol{\omega}_0, \hat{\boldsymbol{\theta}}}. \end{aligned}$$

In matrix form, considering (3)

$$\Delta = \frac{1}{\phi} X^t \operatorname{diag}(\mathbf{y} - \hat{\boldsymbol{\mu}}) \in \mathbb{R}^{(q+1)\times n}. \tag{10}$$

Step 3. For this scheme of perturbation $\mathbf{d}_{\max}$ is the eigenvector correspondent to the greater eigenvalue of

$$\frac{1}{\hat{\phi}} \operatorname{diag}(\mathbf{y} - \hat{\boldsymbol{\mu}})\, X\, (X^t \operatorname{diag}(\hat{\boldsymbol{\mu}}^p)\, X)^{-1}\, X^t \operatorname{diag}(\mathbf{y} - \hat{\boldsymbol{\mu}}). \tag{11}$$

This expression agrees with the results presented by Xie and Wei [27]. In this scenario, an index plot of $\mathbf{d}_{\max}$ will reveal those observations for which a small change produces great changes in the estimators.

3.2. Perturbing one covariable

In this subsection, perturbations on a particular continuous covariable will be considered. We search points, whose values in that covariable exert a great influence on the model likelihood.

Step 1. Let us suppose that the covariable being analyzed is $\mathbf{x}_r$, the r-th column of matrix X will be perturbed by adding a vector $\boldsymbol{\omega}$ scaled by the norm of the column $\mathbf{s}_r$. (Beckman et al. [2]). The perturbed column becomes

$$\mathbf{x}_r^{\omega} = \mathbf{x}_r + \|\mathbf{s}_r\|\boldsymbol{\omega},$$

and $\boldsymbol{\omega} = \mathbf{0} \in R^n$ represents a null perturbation. This scheme will affect the likelihood function only through the estimator of the position parameter θ, the perturbed vector is determined element wise by

$$\theta_i^{\omega} = \sum_{j \neq r}^{q} x_{ij}\beta_j^{\omega} + (x_{ir} + s_r\boldsymbol{\omega})\beta_j^{\omega}, \quad 1 \leq i \leq n. \tag{12}$$

The corresponding perturbed likelihood function is

$$L(\boldsymbol{\beta} / \boldsymbol{\omega}) = \sum_{i=1}^{n}\left(\log c_p(y_i, \phi) + \frac{1}{\phi}(y_i\theta_i^{\omega} - \kappa_p(\theta_i^{\omega}))\right). \tag{13}$$

Step 2. It can be deduced from (12) that

$$\frac{\partial\theta_{\omega i}}{\partial\beta_j} = \begin{cases} x_{ij}, & \text{when } j \neq r, \\ x_{ir} + s_r\omega_i, & \text{when } j = r. \end{cases} \tag{14}$$

Now from (12) and (14),

$$\frac{\partial L_i}{\partial \beta_j} = \begin{cases} \frac{1}{\phi}(y_i - \mu_i^{\omega})\, x_{ij}, & \text{when } j \neq r, \\ \frac{1}{\phi}(y_i - \mu_i^{\omega})(x_{ir} + s_r \omega_i), & \text{when } j = r. \end{cases}$$

Taking derivatives with respect to ω_i, when $j \neq r$

$$\frac{\partial}{\partial \omega_i}\frac{\partial L_i}{\partial \beta_j} = -\frac{1}{\phi}\frac{\partial \mu_i^{\omega}}{\partial \theta_i}\frac{\partial \theta_i}{\partial \omega_i} x_{ij}$$

$$= -\frac{s_r}{\phi}\mu_i^p \beta_{\omega r} x_{ij},$$

and when $j = r$

$$\frac{\partial}{\partial \omega_i}\frac{\partial L_i}{\partial \beta_j} = \frac{1}{\phi}(-\mu_i^p s_r \beta_{\omega r} x_{ij} + y_i s_r - \mu_i s_r)$$

$$= \frac{s_r}{\phi}(y_i - \mu_i - \mu_i^p \beta_{\omega r} x_{ij}).$$

In matrix form,

$$\Delta = \frac{s_r}{\hat{\phi}}\left(\mathbf{u}_r\,(\,\mathbf{y} - \hat{\boldsymbol{\mu}}\,)^t - \hat{\beta}_r X^t \operatorname{diag}(\,\hat{\boldsymbol{\mu}}^p\,)\right), \tag{15}$$

where $\mathbf{u}_r \in \mathbb{R}^{q+1}$ is a vector with 1 in the r-th place and 0 elsewhere. It is a particular case of the expression given for Δ by Thomas and Cook in [24]. Similar results have been reported by Tang et al. in [23] for reproductive dispersion models and by Xie and Wei in [27] for Poisson inverse Gaussian models with equi-dispersion.

Step 3. Viewing (15), $\mathbf{d}_{\max}$ is the eigenvector correspondent to the greater eigenvalue of

$$\frac{s_r}{\hat{\phi}}\left((\,\mathbf{y} - \hat{\boldsymbol{\mu}}\,)\,\mathbf{u}_r^t - \hat{\beta}_r \operatorname{diag}(\,\boldsymbol{\mu}^p\,)\,X\right)(\,X^t \operatorname{diag}(\,\boldsymbol{\mu}^p\,)\,X\,)^{-1}$$

$$\times \left(\mathbf{u}_r\,(\,\mathbf{y} - \hat{\boldsymbol{\mu}}\,)^t - \hat{\beta}_r X^t \operatorname{diag}(\,\boldsymbol{\mu}^p\,)\right).$$

A plot of $\mathbf{d}_{\max}$ versus the index i will show that observations, where a small change in column $\mathbf{x}_r$ generates great changes in the estimators.

3.3. Perturbing the response

Step 1. Following Thomas and Cook [24], in order to perturb the response $\mathbf{y}$, we add a vector $\boldsymbol{\omega}$ scaled by an estimate of the standard deviation of each observation: $s_i = \sqrt{\hat{\phi}\mu_i^p}$. The perturbed vector of observations is $\mathbf{y}^{\omega} = \mathbf{y} + \mathbf{s} \odot \omega$, where $\odot$ is the element wise multiplication; $\boldsymbol{\omega} = 0$ represents no perturbation and the corresponding log-likelihood is given by

$$L(\beta/\boldsymbol{\omega}) = \sum_{i=1}^{n}\left(\log c_p(y_i + s_i\omega_i,\ \phi) + \frac{1}{\phi}\left((y_i + s_i\omega_i)\theta_i - \frac{1}{2-p}((1-p)\theta_i)^{\frac{p-2}{p-1}}\right)\right). \tag{16}$$

Step 2. Differentiating (16) with respect to $\boldsymbol{\beta}$

$$\begin{aligned}\frac{\partial L_i}{\partial \beta_j} &= \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}\\ &= \frac{1}{\phi}\left((y_i + s_i\omega_i)\mu_i^{-p} - \mu_i^{1-p}\right)\mu_i^p x_{ij}\\ &= \frac{1}{\phi}(y_i + s_i\omega_i - \mu_i)\,x_{ij},\end{aligned}$$

and differentiating then with respect to $\boldsymbol{\omega}$

$$\frac{\partial^2 L}{\partial \omega_i \partial \beta_j} = \frac{1}{\phi} s_i x_{ij}.$$

In matrix form and specializing in the *MLE*

$$\Delta = \frac{1}{\hat{\phi}} X^t \operatorname{diag}\left(\hat{\boldsymbol{\mu}}^{-p}\right) S,$$

being $S = \operatorname{diag}(s_i)$. The same result has been reported by Xie and Wei [27] for $s = \phi = 1$.

Step 3. The direction of maximum curvature will be determined by the eigenvector associated with the maximum eigenvalue of the following matrix:

$$S\text{diag}\left(\hat{\boldsymbol{\mu}}^{-p}\right) X\,(X^t\text{diag}\,(\,\hat{\boldsymbol{\mu}}^{\,p}\,)\,X\,)^{-1}\,X^t\text{diag}\left(\hat{\boldsymbol{\mu}}^{-p}\right) S,$$

and will reveal those cases for which a small change in the response produces big changes in $\hat{\boldsymbol{\beta}}$. Again, this is a particular case of the method proposed by Thomas and Cook [24] and also agrees with Tang et al. [23].

3.4. Perturbing the power parameter

Parameter p defines the mean-variance relationship and it is estimated maximizing a profile likelihood curve as described by Dunn and Smyth [6]. If the data set includes points that are overly influential, it may be that the estimator is biased and, consequently, the results obtained would be erroneous. The purpose of this subsection is to define a scheme of perturbations in order to evaluate the sensitivity of maximum likelihood estimators to modifications in p. The perturbation scheme should be defined in such a way that the perturbed parameter $p\omega$ belongs to (1, 2); an option is given by

$$p\omega = (p-1)^{\boldsymbol{\omega}} + 1, \quad \omega \in \mathbb{R}. \tag{17}$$

In this way, for each fixed $p \in (1, 2)$, $p\omega$ can take any value between 1 and 2 with ω varying in $\mathbb{R}$ and with this scheme, all possible pairs $(p,\ p\omega)$ in $(1, 2)\times(1, 2)$ are considered; $\omega = 1$ represents no perturbation.

Step 1. The perturbed log-likelihood is given by

$$L(\boldsymbol{\beta}\,/\,\boldsymbol{\omega}) = \sum_{i=1}^{n}\left(\ln c_{p\omega_i}(y_i,\ \phi) + \frac{1}{\phi}\left(y_i\theta_i - \frac{1}{2-p\omega_i}((1-p\omega_i)\theta_i)^{\frac{p\omega_i-2}{p\omega_i-1}}\right)\right),$$

and the score functions are

$$\frac{\partial L}{\partial \beta_j} = \frac{\partial L}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i - ((1-p\omega_i)\theta_i)^{p\omega_i - 1}\right)x_{ij}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\left(y_i + (-(p-1)^{\omega_i}\theta_i)^{(p-1)\omega_i}\right)x_{ij}.$$

This result agrees with Svetliza and Paula [22]. Differentiating with respect to $\boldsymbol{\omega}$

$$\frac{\partial}{\partial \omega}\left(\frac{\partial L}{\partial \beta_j}\right) = -\frac{1}{\phi}\sum_{i=1}^{n}(\ln(p-1))(-\theta_i(p-1)^{\omega_i})^{(p-1)\omega_i}$$

$$\times (\ln(-\theta_i(p-1)^{\omega_i})+1)(p-1)^{\omega_i}x_{ij},$$

now taking $\omega_i = 1$ and $\hat{\theta}_i = \hat{\mu}_i^{1-p}/(1-p)$,

$$\Delta_{ji} = \frac{1}{\phi}\hat{\mu}_i \ln(p-1)\left(1 - \ln\left(\hat{\mu}_i^{1-p}\right)\right)(p-1)x_{ij},$$

and in matrix form,

$$\Delta = -\frac{1}{\phi}X^t A,$$

where A is a diagonal matrix with elements $a_{ii} = \hat{\mu}_i \ln(p-1)\left(\ln \hat{\mu}_i^{1-p} - 1\right)$ $/(1-p)$. Note that, for extremes values of p, $\Delta_{ji} \overset{p\to 1}{\to} -\infty$ and $\Delta_{ji} \overset{p\to 2}{\to} 0$.

Step 3. The matrix $\Delta^t \ddot{L}^{-1}\Delta$ can be written as

$$A^t X\,(X^t \mathrm{diag}\,(\hat{\boldsymbol{\mu}}^p)\,X)^{-1}\,X^t A, \tag{18}$$

and an index plot of the elements of the eigenvector $\mathbf{d}_{\max}$ will reveal that cases that are influential, when perturbing the mean-variance relationship.

4. Simulation Study

The goal of this section is to explore the effects of perturbing the mean-variance relation, under different settings for p and ϕ. One covariable was generated as $\mathbf{x}_i^t = [1 \quad x_i]$, where $\mathbf{x}$ has uniform distribution in $(0, 1)$, and $\boldsymbol{\mu}$ was calculated applying the inverse link function as follows:

$$\mu_i = g^{-1}(\mathbf{x}_i^t\beta) = -(1-p)(\mathbf{x}_i^t\beta)^{1/(1-p)}.$$

We took $\beta_0 = \beta_1 = -1$ and, for each combination of $p \in \{1.2, 1.5, 1.8\}$ and $\phi \in \{0.01, 0.10\}$, sets of 50 cases were generated as

$$y_i \sim Tw_p(\mu_i, \phi), \quad 1 \le i \le 50. \tag{19}$$

All simulations were carried out in R (R Development Core Team, 2010). The function *r*tweedie( ) from the Tweedie package ([6]) was used to generate observed values and models were fitted by using the Tweedie family option from the statmod package, for the distribution of the response variable.

To disturb p a perturbation was applied on the 90th percentile of $\boldsymbol{\mu}(p_{90})$ and to a point chosen at random (P_{rand}); 500 replics were performed. Two values were chosen for $\boldsymbol{\omega}$ in (17): 0.13 and 7.5, in such a way that, when $p = 1.2$, $p\omega \simeq 1.8$ and vice-versa; when $p = 1.5$, $p\omega \simeq 1.005$ and 1.9. Two benchmarks were used

$$|\text{median}(\mathbf{d}_{\max}) + 3\text{MAD}(\mathbf{d}_{\max})|,$$

$$|\text{median}(\mathbf{d}_{\max}) + 5\text{MAD}(\mathbf{d}_{\max})|.$$

The proportions of replics with P_{90} or P_{rand} greater than the cutting points were calculated and they are shown in Table 1. Also, a graphical representation of these results is given in Figure 1. As can be seen there, when the 90th percentile is disturbed, the effect is almost always detected

by the method, but when the perturbed point is chosen at random, the proportion of cases detected is considerably lower. As expected, greater dispersion is associated with less sensibility. On the other hand, when $p = 1.2$ or 1.8 (near the limits of the allowed interval (1, 2)), the method is less sensible than when $p = 1.5$.

Table 1. Percentage of times P_{90} or P_{rand} were greater than the cutting point

		P_{90}	P_{90}	P_{rand}	P_{rand}
p	ϕ	3 MAD	5 MAD	3 MAD	5 MAD
1.2	0.01	684 (85.5%)	775 (96.9%)	612 (75.5%)	763 (95.4%)
	0.10	624 (78.0%)	774 (96.8%)	518 (64.8%)	757 (94.6%)
1.5	0.01	680 (85.0%)	747 (93.4%)	606 (75.8%)	751 (93.9%)
	0.10	565 (70.6%)	729 (91.1%)	495 (61.9%)	701 (87.6%)
1.8	0.01	735 (91.9%)	781 (97.6%)	677 (84.6%)	766 (95.8%)
	0.10	588 (73.5%)	726 (90.8%)	526 (65.8%)	677 (84.6%)

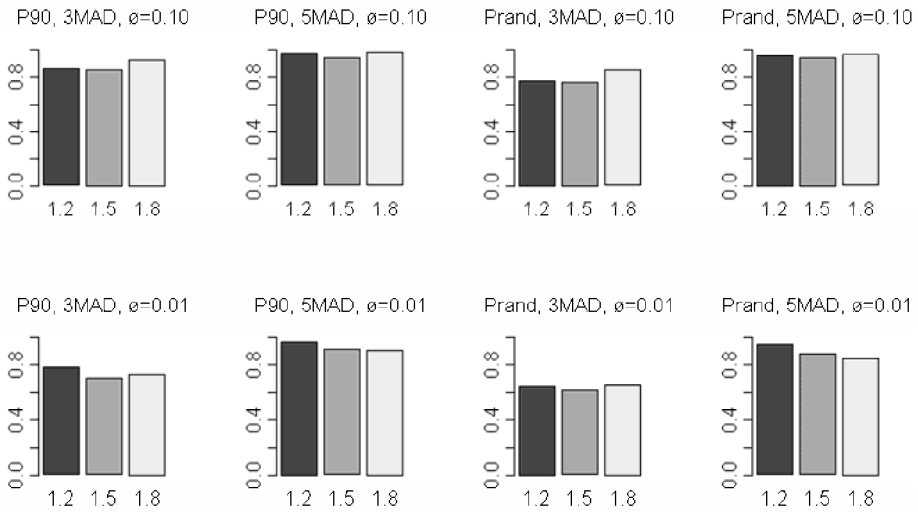

Figure 1. Comparative bars for different cut points, perturbations, densities, and dispersions.

5. An Application to Real Data

We present in this section an application of the local influence approach to data from the "Encuesta Nacional de Gastos de Hogares", carried out each ten years by the "Instituto Nacional de Estadísticas *y* Censos" in Argentina; we considered a data set constituted by 2869 homes from Buenos Aires region in a given week in 1998. This survey registers home expenditures in recreation and leisure activity, classified in various concepts that were registered as shown in Table 2: each column represents a different concept (cinema, travels, etc.) and N is the number of concepts with positive expenditures. Total expenditure (*TE*) was defined as

$$y_i = \sum_{j=1}^{N_i} x_{ij}, \quad 1 \le i \le n,$$

where y_i is the value of *TE* in the i-th home, x_{ij} is the expenditure in concept j, and N_i is the number of concepts with null expenditure. A total of 710 homes (25%) had no expenditures in any concept giving $N_i = 0$ and $y_i = 0$. For the remaining homes, the mean value of *TE* was \$68.46, and the maximum was \$574.00.

Table 2. Data matrix format

	Concept A	Concept B	Concept C	Concept D	Total	N
Home 1	\$28,23	\$32,14	\$50,00	\$26,30	\$136,67	4
Home 2	\$25,00	\$0,00	\$69,00	\$0,00	\$94,00	3
Home 3	\$0,00	\$0,00	\$0,00	\$0,00	\$0,00	0
Home 4	\$78,25	\$15,60	\$0,00	\$14,25	\$108,10	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Home 2869	\$17,00	\$0,00	\$15,00	\$0,00	\$32,00	2

A generalized linear model was fitted to this data by Alegre et al. [1], considering the response *TE* as a compound-Poisson variable for which, the parameter p was estimated via profile likelihood (see [6]), the maximum was reached at $p = 1.4$.

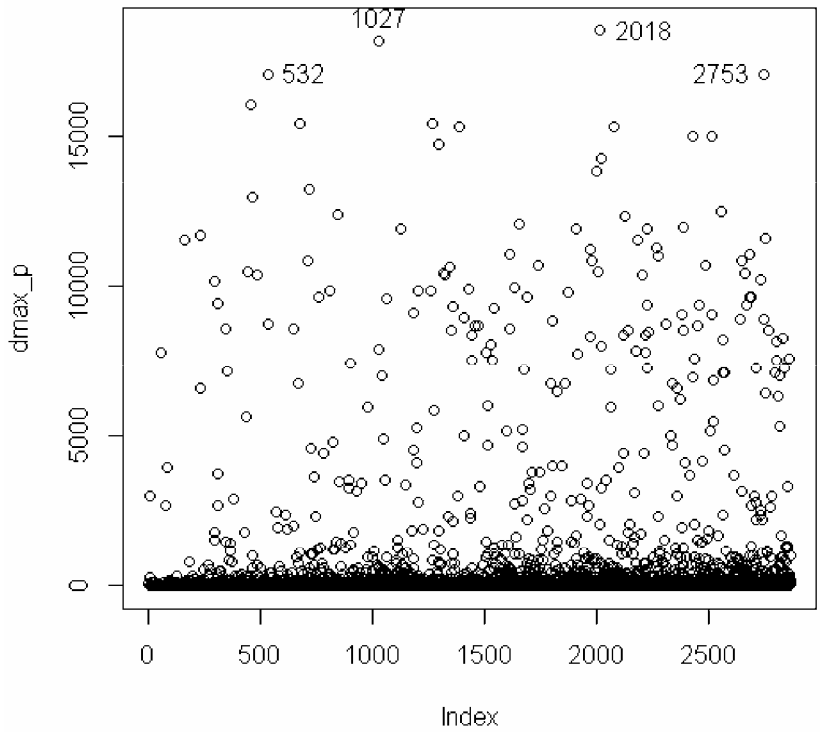

Figure 2. Index plot of $\mathbf{d}_{\max}$ when perturbing the mean-variance relation.

Five covariables were selected in the final model: head of household age (*AGE*), head of household educational level (*EDLEV*), number of members living in that house (*NUM*), car propriety (*CAR*), and family income quintil (*INC_Q*). Perturbations to the mean-variance relation were analyzed, following the scheme given in Subsection 3.3. The index plot can be seen in Figure 2, clearly, the four most influential points correspond to observations 532, 1027, 2018, and 2753. They all were young people, with high income, high educational level, and high expenditure in recreation and leisure.

6. Concluding Remarks and Future Research

We have discussed in this article applications of influence diagnostic in compound-Poisson regression models. Four perturbation schemes were considered, three of them known for other models: perturbing cases, perturbing covariables, and perturbing the response, and the fourth one that perturbs the mean-variance relation is inherent to compound-Poisson models. The corresponding matrices, whose greater eigenvector will detect influence points, were derived for these schemes. A study carried out by simulations evidenced lightly more sensibility when the parameter p of the compound-Poisson density approaches 1 or 2. As expected, for increasing dispersion, there is less sensibility. A data set about home expenditures was analyzed and influential points were detected and characterized.

References

[1] P. Alegre, N. Liseras and L. Ricci, Una aplicación de los modelos Tweedie a las decisiones económicas de los hogares, VII Congreso Latinoamericano de Sociedades de Estadística, Rosario, Argentina, 7 (2006), 16-17.

[2] R. J. Beckman, C. J. Natsheim and R. Dennis Cook, Diagnostics for mixed-models analysis of variance, Technometrics 29 (1987), 413-426.

[3] S. G. Candy, Modelling catch and effort data using generalized linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects, CCAMLR Science 11 (2004), 59-80.

[4] R. D. Cook, Assesment of local influence, Journal of the Royal Statistical Society B 48 (1986), 133-169.

[5] P. K. Dunn, Occurrence and quantity of precipitation can be modelled simultaneously, International Journal of Climatology 24 (2004), 1231-1239.

[6] P. K. Dunn and G. K. Smyth, Series evaluation of Tweedie exponential dispersion model densities, Statistics and Computing 15 (2005), 267-280.

[7] P. K. Dunn and G. K. Smyth, Evaluation of Tweedie exponential dispersion model densities by Fourier inversion, Statistics and Computing 18 (2008), 73-86.

[8] Y.-Z. Fu, N.-S. Tang and X. Chen, Local influence analysis of nonlinear structural equation models with nonignorable missing outcomes from reproductive dispersion models, Computational Statistics and Data Analysis 53 (2009), 3671-3684.

[9] M. Galea, H. Bolfarine and F. Vilca, Local influence in comparative calibration models under elliptical t-distributions, Biometrical Journal 47 (2005), 691-706.

[10] J. Hardin and J. Hilbe, Generalized Linear Models and Extensions, Stata Press, 2001.

[11] B. Jorgensen, The Theory of Exponential Dispersion Models and Analysis of Deviance, Volume 51, Monografías de Matemática, IMPA, Rio de Janeiro, Brasil, 1992.

[12] B. Jorgensen, The Theory of Dispersion Models, Chapman and Hall, 1997.

[13] B. Jørgensen, J. R. Martínez and V. Vinogradov, Domains of attraction to Tweedie distributions, Lithuanian Mathematical Journal 49 (2009), 399-425.

[14] A. J. Lawrence, Regression transformation diagnostics using local influence, Journal of the American Statistical Association 84 (1988), 125-141.

[15] G. Molenberghs, G. Verbeke, H. Thijs, E. Lesaffre and M. G. Kenward, Influence analysis to assess sensitivity of the drop out process, Computational Statistics and Data Analysis 37 (2001), 93-113.

[16] J. A. Nelder and R. W. M. Wedderburn, Generalized linear models, Journal of the Royal Statistical Society, Series A 135 (1992), 370-384.

[17] E. M. M. Ortega, H. Bolfarine and G. A. Paula, Influence diagnostic in generalized log-gamma regression models, Computational Statistics and Data Analysis 42 (2003), 165-186.

[18] F. Osorio, G. A. Paula and M. Galea, Assesment of local influence in elliptical linear models with longitudinal structure, Computational Statistics and Data Analysis 51 (2007), 4354-4368.

[19] G. A. Paula, Assessing local influence in restricted regression models, Computational Statistics and Data Analysis 16 (1993), 73-79.

[20] H. Shono, Application of the Tweedie distribution to zero-catch data in CPUE analysis, Fisheries Research 93 (2008), 154-162.

[21] G. K. Smyth and B. Jorgensen, Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling, Astin Bulletin 32 (2002), 143-157.

[22] C. F. Svetliza and G. A. Paula, Diagnostics in nonlinear negative binomial models, Communications in Statistics 32 (2003), 1227-1250.

[23] N.-S. Tang, B.-Ch. Wei and X.-R. Wang, Local influence in nonlinear reproductive dispersion models, Communications in Statistics, Theory and Methods 30 (2001), 435-449.

[24] W. Thomas and R. D. Cook, Assesing influence on regression coefficients in generalized linear models, Biometrika 76 (1989), 741-749.

[25] C. H. Tsai and X. Wu, Assessing local influence in linear regression models with first order autorregressive or heteroscedastic errors structure, Statistics and Probability Letters 14 (1992), 247-252.

[26] M. Tweedie, An index which distinguishes between some important exponential families, Statistics: Applications and new directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Series A 135 (1984), 579-604.

[27] F.-Ch. Xie and B.-Ch. Wei, Influence analysis in Poisson inverse Gaussian regression models based on the em algorithm, Metrika 67 (2008), 49-72.

■