REVISTA DE ESTATÍSTICA
3º QUADRIMESTRE DE 2001

A NEW LOCAL INFLUENCE MEASURE IN GENERALIZED LINEAR MODELS

UMA NOVA MEDIDA DE INFLUÊNCIA LOCAL EM MODELOS LINEARES GENERALIZADOS

Autores: Yenis Marisel González Mora
- Department of Statistics, Research Operation and Computation. Faculty of Mathematics. University of La Laguna. Tenerife

M. Mercedes Suárez Rancel
- Department of Statistics, Research Operation and Computation. Faculty of Mathematics. University of La Laguna. Tenerife

ABSTRACT:

- This paper investigates the local influence assessment in Poisson generalized linear models. The assessment of local influence in generalized linear models is studied by the approach of Cook (1986). So he only deals with the local influence on the regression coefficients which are not resistant to masking and swamping effects. Suárez and González (2000) propose a new measure to detect locally influential data under perturbations to variance. Based on this measure, in this paper we propose a locally influential measure to mitigate these difficulties. We demonstrate the need of this measure. An example shows the effectiveness of the proposed method.

KEY-WORDS:

- *Generalized linear model, Masking and Swamping, Local influence.*

***RESUMO*:**

- Este artigo procede à avaliação da influência local em modelos lineares generalizados de Poisson. A avaliação da influência local em modelos lineares generalizados é estudada pela abordagem de Cook (1986). Este autor lida apenas com o caso da influência local em coeficientes de regressão que não são robustos a efeitos masking e swamping. Suárez and González (2000) propõem uma nova medida para detectar dados localmente influentes sob perturbações da variância. Neste artigo, propomos uma medida de influência local baseada nessa medida, que minora estas dificuldades e demonstramos a necessidade de tal medida. Um exemplo ilustra a eficácia do método proposto.

***PALAVRAS-CHAVE*:**

- *Modelos Lineares Generalizados, Masking e Swamping, influência local.*

1. INTRODUCTION

We assume the observations consist of a vector y of n independent responses from the exponential family. We study a Poisson generalized linear model, which is a particularly case of this family.

The idea of influence assessment is to monitor the sensitivity of statistical analysis subject to minor changes in the model. The works of some authors (Lawrence, 1988, Peña and Yohai, 1995) indicate that one of the attractions of the local influence concept is that it assesses the effect of joint perturbations on the data cases more easily than global influence measures Suárez and González (2000). Thus in a local sense, frequently, the results are free from masking effects that present difficulties for individual case-deletion methods.

The purpose of this study is to gain additional insight on global influence regarding the local influence analysis and its implications.

In the next section we give the general idea of generalized linear models. Section 3 we describe a general idea of local influence. Section 4 shows that local–influence analysis of perturbations of the variance is similar to the usual regression diagnostic based on Hadi's measure for detecting an influential subset. In Section 5 we extend the 'Suárez and Glez' measure to Poisson generalized linear model. In Section 6 we give a Lawrence's transformation measure. And Section 6 provides an illustrative example.

2. A POISSON GENERALIZED LINEAR MODEL

2.1. RESPONSE DISTRIBUTION

We assume the observations consist of a vector y of n independent responses from the exponential family

$$f(y;\theta;\phi) = \exp\{[\, y\theta - b(\theta)\ ]/\, a(\phi) + c(y;\phi)\}\, y = 0,1,2,...$$

with $\theta_i = g(\eta_i)$, $\eta_i = x_i'\beta$, where x is an n×p matrix of covariates, is a p-dimensional column vector of unknown parameters, and a(.),b(.),c(.) are known functions. The dispersion parameter ϕ is usually regarded as nuisance parameter.

Then the mean and variance of y can be written by:

$$E[y] = b'(\theta) = \mu$$
$$Var[y] = b''(\theta)a(\phi) = b''(\theta)\,\phi / w,$$

where the primes denote derivates with respect to θ.

Applying this in Poisson distribution we have:

$$f(y;\theta;\phi) = \exp\{(y \ln \lambda - \lambda) - \ln y!\} y = 0,1,2,... \quad (1)$$

$\phi = 1$, $b(\theta) = \lambda$, $c(y, \phi) = -\ln y!$

$E[Y] = \lambda$

$Var[Y] = \lambda$

2.2. LINK FUNCTION

The mean μ_i of the response in the i-th observation is related to a linear predictor through a monotonic differentiable link function g

$$\eta_i = g(\mu_i) = x_i'\beta$$

Here, x_i is a fixed known vector of explanatory variables, and β is a vector of unknown parameters.

In classical linear models they are identical, and the identity link is sensible in the sense that both η and μ can take any value on the real line. However, when are dealing with counts and the distribution is Poisson, we must have $\mu>0$, so that the identity link is less attractive in part because may then be negative. Models based on independence of probabilities associated with the different classifications of cross-classified data lead naturally to considering multiplicative effects, and this is expressed by the log link, $\eta = \log \mu$ with its inverse $\mu = e^{\eta}$ .Therefore the link function for Poisson (λ) models is the log link $\eta = \log \lambda$.

2.3. MAXIMUM LIKELIHOOD FITTING

An iterative methods are required to solve the normals equations:

$$X^t s = X^t (y - \hat{y}) = 0,$$

And these lead to the iterative scheme:

$$\beta^{t+1} = \beta^t - H^{-1} s,$$

where H is the hessian matrix and s is the gradient vector of the log-likelihood function. That is:

$$s=[s_j]=[\partial L/\partial \beta_j] \text{ and } H=[h_{ij}]= [\partial^2 L/\partial \beta_i \partial \beta_j].$$

The estimated covariance matrix of the parameter estimator is given by

$$\Sigma = -H^{-1}$$

In our case, $\Sigma = (X^t V X)^{-1}$, where V is a diagonal matrix with the variance λ_i of the response Y(~ Poisson(λ)) in the i-th observation.

2.4. GOODNESS OF FIT

Two statistics that are helpful in assesing the goodness of fit of a given generalized linear model are the scaled deviance and Pearson's chi-square statistic.

The scaled deviance is defined by

$$D(y,\mu) = 2\{l(X\hat{\beta}; y) - l(\hat{\theta}; y)\},$$

where $l(\hat{\theta}; y)$ refers to the maximum of the log-likelihood function based on fitting each exactly. Therefore, the deviance in Poisson models can be written as:

$$2\{ \sum y \ln(y/\lambda) + \sum (y-\lambda)\}$$

And finally, the Pearson'chi-square statistic is defined as:

$$\chi^2 = \Sigma(y_i-\mu_i)^2 /V(\mu_i)$$

where $\hat{\mu}_i = b^{(1)}\left(g\left(x_i\hat{\beta}\right)\right)$ and $Var(\hat{\mu}_i) = b^{(2)}\left(g\left(x_i\hat{\beta}\right)\right)$.is the estimated variance function for the distribution concerned..

3. LOCAL AND DELETION INFLUENCE

We shall give a brief review of Cook's local influence approach in this section to provide some help for understanding the new formulation which will be defined in Section4.

Consider the Poisson generalized linear model defined in Section 1

$$f(y;\theta;\phi) = \exp\{(y\ln\lambda - \lambda) - \ln y!\} y = 0,1,2,...$$

with $\theta_i = ln\ \lambda$, $\eta_i = X_i\beta$. Let $\hat{\beta}$ be the estimated parameter vector of β (MLE).

Many measures have been suggested to assess the influence of observations. Cook(1986) considers a general version of Cook's distance

$$D_i \equiv \frac{\| \hat{Y} - \hat{Y}_{(i)} \|^2}{k\sigma^2} ,$$

where $\hat{Y}, \hat{Y}_{(i)}$ are the n×1 vectors of the fitted values based on the full data and the data without i-th case, respectively, and k is the dimension of β.

He investigated

$$D_i(w) = \frac{\| \hat{Y} - \hat{Y}_{(w)} \|^2}{k\sigma^2} ,$$

where $\hat{Y}_{(w)}$ is the vector of fitted values when the i-th case has weight w and the remaining cases have weight 1.

These ideas have been extended to general cases . In generalized linear models the generalized Cook's distance(McCullagh and Nelder(1989)) is defined by

$$LD_i = \left(\hat{\beta}_{(i)} - \hat{\beta}\right)'(x'Wx)\left(\hat{\beta}_{(i)} - \hat{\beta}\right)/\hat{\phi}$$

as a measure of the effect of the i-th datum point on the parameter estimates, where $W = diag\{W_i\}$, $W_i = b^{(2)}(\hat{\theta}_i)\left[g^{(1)}(\hat{\eta}_i)\right]^2$, $\hat{\beta}_i$ denotes the estimates when that point is omitted and the superscript (2) denotes the 2nd derivative of the function.

An approximate measure of leverage is given by the diagonal element h_i of the projection matrix

$$H = W^{1/2}x(x'Wx)^{-1}x'W^{1/2}$$

Cook(1986) developed a general technique for the assement of local influence. Sensivity of the analysis was assesed though the normal curvature of the likelihood displacement surface when minor perturbations were introduced in the postulated model. Further extentions to generalized linear model were given by Thomas and Cook(1989). The likelihood displacement takes the form

$$LD(w) = 2\left[L(\hat{\beta}) - L(\hat{\beta}_w)\right] ,$$

where $L(\hat{\beta})$ is the likelihood for β and write $\hat{\beta}_w$ for the MLE from the perturbed model. The vector of the values w and LD(w) from the surface of interest as w varies. The direction h_{max} of the maximum curvature of the likelihood displacement surface in the postulated model(where $w = w_0$) indicates the greatest local sensitivity against perturbations. The direction of maximum curvature is used as the main diagnostic tool in the local influence method. But this measure has some practical and theoretical difficulties. For example, computability of the maximum curvature is restricted to the linear regression model; there is also a lack of invariance of the curvature under reparametrisation of the perturbation scheme; and lack of definition of the parameters.

4. DIAGNOSTIC IN POISSON GENERALIZED LINEAR MODEL

We show that local-influence analysis of perturbations of the variance is similar to Hadi's measure for detecting an influential subset. To avoid the difficulties defined in Section 3, Suárez and González(2000) suggest an alternative likelihood

displacement. Based on this likelihood displacement, we propose a quasilikelihood displacement to mitigate the swamping and masking effect:

$$LD_{(i)}(w_i) = -2\left[L(\hat{\beta}) - L\ (\hat{\beta}_w \mid w) \right] + \left[\text{var}\left(\hat{\mu}_i \right) - \text{var}\left(\hat{\mu}_{w_i} \right) \right]$$

where $\hat{\mu}_i = b^{(1)}\left(g\left(x_i \hat{\beta}\right)\right)$ and $Var(\hat{\mu}_i) = b^{(2)}\left(g\left(x_i \hat{\beta}\right)\right)$.

Therefore , the aim of the present paper is give a measure to mitigate these difficulties in Poisson generalized linear model.

Applying this displacement to Poisson model we obtain the slope:

$$LD^{**}{}_i(w_i) = \hat{y}_i - y_i - \frac{h_i{}^2}{m_i\ \hat{\lambda}_i},$$

Where h_i are the diagonal elements of the projection matrix H defined is section 2, $\hat{y}_i$ are the predicted values of response variable, m_i is the number of observations in the i-th covariate and λ_i are the diagonal element of the variance matrix V.

However, in Poisson generalized linear model we have no constant variance, therefore we use Box and Cox (1964) transformation in model (1) to mitigate this problem.

Then, we have:

$$LD^{**}{}_i(w_i) = \hat{y}_i - \sqrt{y_i} - \frac{h_i{}^2}{m_i\ \hat{\lambda}_i},$$

5. TRASFORMATION DIAGNOSTIC IN POISSON MODEL.

The aim of diagnostic in regression analysis is to appropriateness of the assumptions made in fitting a regression model to the data. Lawrence(1988) obtains diagnostics for the estimated regression parameter of the Box and Cox transformation of the response variable in the linear model:

Let $Y^{(\lambda)}$ (Poisson model defined in section1) be :

$$f(y;\theta;\phi) = \exp\{(y \ln \lambda - \lambda) - \ln y!\} y = 0,1,2,...$$

$$z^{(\lambda)} = y^{(\lambda)} / J(\lambda)^{1/n}, \text{ where}$$
$$J(\lambda) = \prod_{i=1}^{n} dy_i^{(\lambda)} / dy_i,$$

and we define z' as the derivate of $z^{(\lambda)}$ with respect to λ.

For the Box-Cox power family transformation we have:

$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda}$ and $z^{(\lambda)} = \frac{y^{\lambda} - 1}{\dot{y}^{\lambda-1}\lambda}$, where $\dot{y}$ is the geometric mean of (y_1 ,y_2 , ... ,y_n).

The variance in the transformed model is slightly perturbed, in particular, it provides to determining those cases of the data that have strongest general influence on the estimated transformations parameters.

Now the variance of the model become:

$$\text{Var}(e) = \sigma^2 W^{-1},$$

where W is a diagonal matrix of perturbations.

Lawrence proposes a measure to assess the effect of joint perturbations on the data cases:

$$LD^* = l_{max_i} = r_i + r'_i / \sum^{n} [\{r_j r'_j\}_2]_{1/2}$$

i= 1,...,n; where r_i and r_i' are the ith residuals from the regression of z and z' on the column of X, respectively.

The diagnostic then arises from local changes to the transformation parameter estimate caused by small perturbations; the case direction in which small perturbations have the greatest effect is the main diagnostic quantity.

6. COMPARISIONS OF LOCAL INFLUENCE MEASURES WITH A POISSON DATA.

We now compare the various local influence in the contex of a Poisson data set.

The data arise from a 5*4 factorial design (Maxwell 1961) with four replications at each factor level. The responses are the number of boys with disturbed dreams, and the factors are Rating (four levels) and Age groups (five levels).

The following model is fit to the data.

$$\text{Ln Dream} = \beta_0 + \beta_1 \text{ Rating} + \beta_2 \text{ Age} + \in,$$

we have n=20 and k=3 in these data. On one hand, we apply LD,LD^*,LD^{**} and the five largest values are in Table 1.

Table 1. Maxwell(1961) data: Five largest values based on LD(Cook's distance), and LD^* (Lawrance's measure), LD^{**} (New measure)

Case	LD	Case	LD^*	Case	LD^{**}
20	4.031	15	0.379	20	19.927
15	1.920	20	0.339	19	17.378
2	1.649	2	0.3153	18	15.152
16	1.486	19	0.304	15	13.982
19	1.078	10	0.301	17	13.392

As can be seen from Table1, observations 20 is more local influential according to LD and LD^{**}.

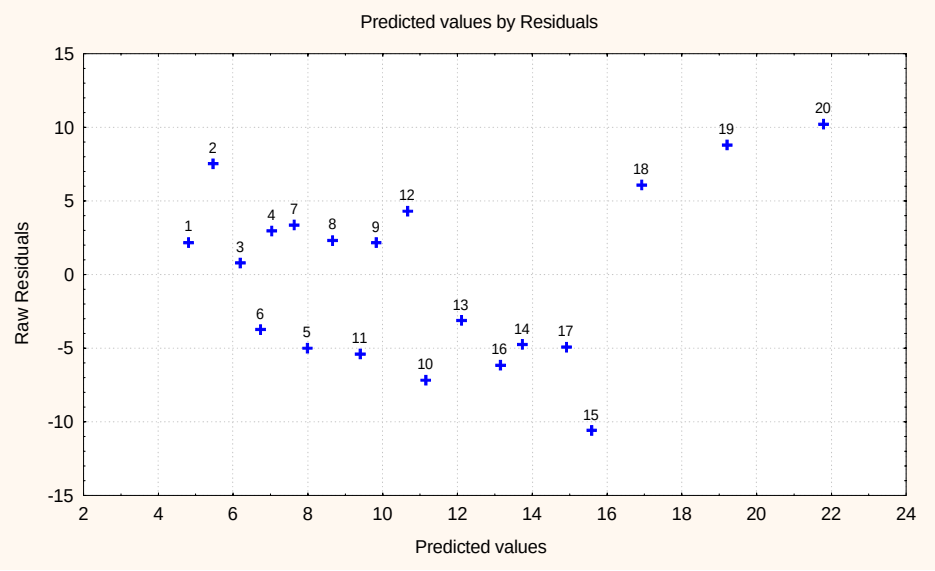

FIG. 1: A plot of Residuals versus Predicted values.

Cases 18,19,20 seem out of line with the rest of the plot.As we can see that LD, LD^* provide a useful method for investigating influence but they have some therorical and practical difficulties like swamping and masking effects.As we can observe in

table 1 they are no resistant to these effects.Also,we can observe that the stranges observations (cases:20,18,19) are detected by LD^{**} and this measure hasn't these problems .

REFERENCES

BILLOR and LOYNES, R. M. (1993), Local influence:A new approach, Communication.Stat.Theory Methods,22,1595-1611.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations,J_R.Stat Soc B,26,211-246.

COOK, R. D. (1977). Detection of influential observations in linear regression.Technometrics,19,15-18.

COOK, R. D. (1986).Assement of local influence,Journal of the Royal Statistical Society SeriesB-Methodological,48,133-169.

HADI, A. S. (1992).A new measure of overall potential influence in linear regression.computational Statistics&Data Analysis 14,1-27.

LAWRENCE, A. J. (1988),Regression transformation diagnostic using local influence, Journal of the American Statistical Association,86,1067-1072.

MAXWELL, J. A. (1961).Analysing qualitative data.London:Methuen.

MACCULLAGH, P., and NELDER, J. A. (1989),Generalized linar models.London:Chapman and Hall.J.A

NELDER, J. A. and WEDDERBURN, W. M. (1972).Generalized linear model, Journal of the Royal Statistical Society,A135,370-384.

PEÑA, D. and YOHAI, V. J. (1995).The detection of influential subsets in linear regression by using an influence matriz, Journal of the Royal Statistical Society SeriesB-Methodological,57,145-156.

SUÁREZ, M. .M. and GONZÁLEZ, M. A. (1996).Medidas basadas en influencia local.Cuadernos de Bioestadística y sus aplicaciones Informáticas, vol 14,5-17.

SUÁREZ, M. .M. and GONZÁLEZ, M. A. (2000).A connection between local and deletion influence, Sankhya:The Indian Journal of Statistics 2000, 62,series A,pt1,144-149.

SUÁREZ, M. .M. and GONZÁLEZ, M. A. (2000).Local and deletion diagnostic, Test, vol 9, No 2,345-352.

THOMAS, W and COOK, R. D. (1989). Assesing influence on regression-coefficients in generalized linear models, Biometrika,76,741-749.