Malaysian Journal of Science **18**: 55-58 (1999)

Least Median of Deviance - an alternative to Maximum Likelihood Estimation in Generalised Linear Models. Application to linear logistion regression

Nor Aishah Hamzali

Department of Mathematics, University of Malaya, 50603 Kuala Lumpur, Malaysia

Abstract. An alternative method of parameter estimation in Generalised Linear Models in the presence of outlying observations is discussed. The method, Least Median of Deviance, is an extension of the exact least median of squares for the linear regression model. A numerical example with application to logistic regression is presented.

Abstrak. Satu kaedah alternatif penganggar parameter di Model Linear Teritlak apabila wujud titik-titik terpencil dibincangkan. Kaedah Median Deviaus (sisihan) Terkecil adalah lanjutan daripada kaedah Median Kuasadua Terkecil bagi model linear regressi. Satu contoh berangka dengan penggunaan kepada regressi logistik dibentangkan.

Introduction

In this paper, we examine the robust estimation of β in generalised linear models (GLMs) [1] when the conditional density of YIX takes the form of

$$f(y_i|\boldsymbol{x}_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi)\right\}$$

where b and c are known function and β is related to θ_i via the relationship

$$E(y_i) = \mu_i(\theta_i)$$

and $g(\mu_i) = \alpha_i\beta = \eta_I$,

where g is a differentiable function known as a *link function.*

The most commonly used method in estimating the unknown parameter β is the maximum likelihood estimation (MLE) in which β_{MLE} maximises over β, the sum of individual log-likelihood functions. However, the MLE is not robust in that it is subject to influence by outliers. In order to provide some protection against small subsets of outlying observations, analysts have made important extensions of the diagnostic as well as robust regression approaches in both linear and non-linear models.

The proposed robust technique studied here is an extension of the Least Median of Squares [2,3] in GLMs, called the Least Median of Deviance (LMD). The LMD estimate is a minimax estimate which minimises the qth ordered deviance for a given data set where $q = [(n - p - 1)/2] + (p + 1)$ and p is the number of unknown parameters to be estimated.

On the theory of minimax (LMD) estimation for GLMs

The minimax estimation for GLMs problem is to find βj to minimize the maximum $d_{[k]}(\beta)$ where d_i denotes the ith deviance of observation y_i from the fitted model and $[t]$ denotes the integer value of t.

Suppose $n = p + 1$. Theorem 1 shows that the minimax solution is the exact solution when $d_i(\beta)$, $(i = 1, 2, \ldots , p + 1)$ are all equal.

Theorem1. *Suppose that g and b' are both strictly monotone. Suppose that $\beta \in \boldsymbol{R}^p$ and that rank $(X) = p$.*

Define

$$d_j(\theta) \quad = \quad 2\{b(\theta) - y_j\theta - b(\hat{\theta}_j) + y_j\hat{\theta}_j\}$$

where

$$\theta_j = \theta_j(\beta) \quad = \quad b'^{-1}(g^{-1}(\mathbf{x}_j^T\beta)).$$

and θj is the mle based in the jth observation alone. Then

(i) there ezists β such that $di(\theta i)$, $d_2(\theta_2)$, ..., $d_{p+1}(\theta_{p+1})$ are all equal,
(ii) the value of β that minimises $max_{1 \leq j \leq p+1} d_j(\theta j)$ is such that

$$d_1(\theta_1) = d_2(\theta_2) = \ldots = d_{p+1}(\Theta_{p+1})$$

The proof of the existence of the minimax LMD estimate can be found in [4].

Remarks. The least median of deviance estimate (LMD) is a special case of a more general estimate called the least quantile of deviance estimate (LQD). In the case of a normal linear regression, the LMD is LMS which is a special case of the least quantile of squares (LQS) [3]. Because an LQD estimate minimizes the *k*th smallest ($k > p$) deviance residual for a given data set, it must minimize the maximum deviance for some *k* element subset of the data. Thus the *k*th LQD estimate must be the minimax deviance fit to that *k* element subset. In principle, all LQD estimates in any model can be calculated by exhaustively searching over subsets of the data of a given size and computing the minimax solution for each subset. Unfortunately, as the number of sample sizes *n* and the number of parameters *p* increases, the computation involved are often infeasible because of the large number of subsets that would have to be considered. In practice, only some percentage of all possible subsets (chosen at random) will be looked at if C^n_{p+1} is large.

A numerical example

We will consider a numerical example to illustrate the need for a robust alternative to the MLE criterion. A complete analysis of the data set is outside the scope of this section as the purpose of this study being to contrast MLE approach with the proposed LMD method of estimation.

A multiple logistic regression example: Vaso constriction of the skin. The data in Table 1, given by Finney (1947, p. 322) consist of 39 binary responses (*y*) denoting the presence (1) or absence (0) of vaso-constriction of the skin of the digits after inspiration of a volume of air *V* at the inspiration rate *R*. A dose-response relationship between the explanatory variables and the dichotomous outcome is the basis for the proposed model.

Table 1. Listing of Finney's data on vaso constriction in the skin of the digits. The binary response y indicates the occurence (1) or nonocccurence (0) of vaso constriction

Volume	Rate	Response	Volume	Rate	Response
3.7	0.825	1	1.8	1.8	1
3.5	1.09	1	0.4	2.0	0
1.25	2.5	1	0.95	1.36	0
0.75	1.5	1	1.35	1.35	0
0.8	3.2	1	1.5	1.36	0
0.7	3.5	1	1.6	1.78	1
0.6	0.75	0	0.6	1.5	0
1.1	1.7	0	1.8	1.5	1
0.9	0.75	0	0.95	1.9	0
0.9	0.45	0	1.9	0.95	1
0.8	0.57	0	1.6	0.4	0
0.55	2.75	0	2.7	0.75	1
0.6	3.0	0	2.35	0.03	0
1.4	2.33	1	1.1	1.83	0
0.75	3.75	1	1.1	2.2	1
2.3	1.64	1	1.2	2.0	1
3.2	1.6	1	0.8	3.33	1
0.85	1.415	1	0.95	1.9	0
1.7	1.06	0	0.75	1.9	0
			1.3	1.625	1

Concluding remarks

The data obtained were repeated measurement on three individual subjects, the numbers of observation per subject being 9, 8 and 22. Finney found no evidence of intersubject variability and was satisfied to treat the data as 39 independent observations. The model under consideration regards the binary outcome y as a Bernoulli variable with parameter π where π is related to the volume and rate of air inspired via the relationship

$$\begin{aligned} logit(\pi) &= \log\{\pi/(1-\pi)\} \\ &= \beta_0 + \beta_1 \log V + \beta_2 \log R. \end{aligned} \qquad (2)$$

Assuming that the model is correctly specified, the MLE fit to this data set yields

$$logit(\pi) = -2.863 + 4.538\log V + 5.122\log R \qquad (3)$$

By using the diagnostic case deletion, Pregibon [5] showed that observations 4 and 18 individually have an enormous effect on the estimated coefficients.

The proposed approximate LMD fit to this data gives

$$logit(\pi) = -34.506 + 47.001\log V + 42.839\log R \qquad (4)$$

The estimates obtained by the approximate LMD (with $m = 3500$ subsamples) are clearly very different from the MLE. This drastic change is due to the fact that the negative responses are very nearly separated from the positive responses by a straight line in the $\log V$, $\log R$ plane [6]. The 4th and 18th observations contributes the most in (3) *i.e* the rise of the logistic surface.

The minimax LMD fit does reveal that observations 4 and 18 are clearly inconsistent with the majority of the observations and this is captured in the residual plot of Figure 2(b). The approximate LMD seems to fit the majority of the data well. Notice that observations 29,31 and 39 are also quite distant from most of the observations.

We note that because the coefficient estimates of $\log V$ and $\log R$ in [5] are similar, this may suggest the possibility of considering RV as a variable.

Concluding remarks

The LMD is the analogue of the LMS in the normal linear regression model, which minimises the maximum deviance of the 'half samples'. Even though this method has a high breakdown point (which will be discussed elsewhere) in many cases, the difficulty of the LMD is that, in general, it is computationally expensive. More work will be required to improve the algorithm involved.

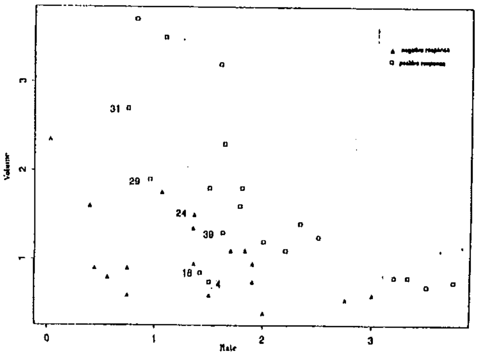


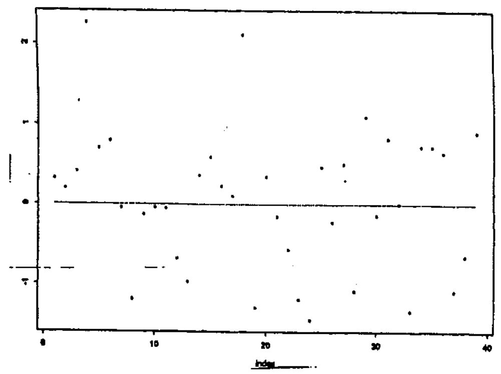
Figure 1. Scatter plot of the vaso constriction of the skin.

Acknowledgements

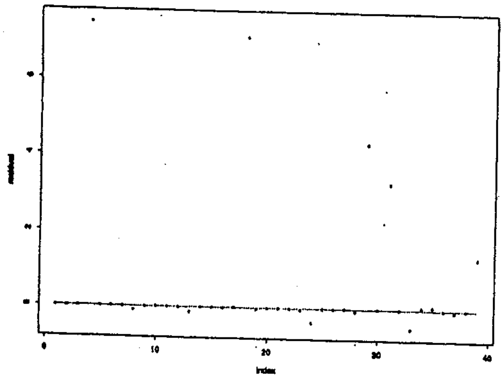
The author would like to thank Professor Peter Green and Dr. David Green for their advice and supervision during the course of this study.

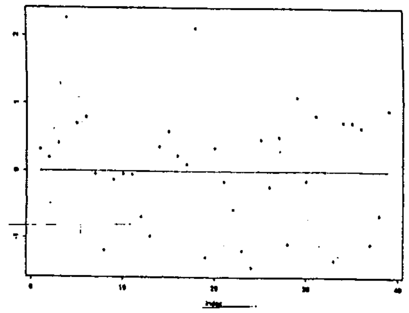
References

1 Nelder, J. A. and Wedderburn, R.W.M. (1972). *J. R. Stat. Soc. A* **135**: 370.
2 Rosseeuw, P.J. (1984). *J. Am. Stat. Assoc.* **79**: 871.
3 Stromberg, A.J. (1993). *SIAM J. Sci. Comput.* **14**: 1289.
4 Hamzah, N. and Green, D. (1997. *J. Statsci.* In press.
5 Pregibon, D. (1982). *Biometrics* **38**: 485.
6 Morgenthaler, S. (1992). *Biometrika* **79**: 747.



Vaso constriction of the skin: Residual plot from MLE fit.





Vaso constriction of the skin: Residual plot from MLE fit.

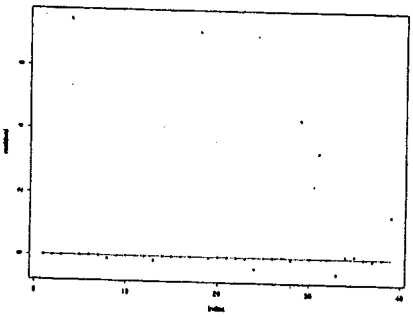


Figure 2. Residual plot from LMD fit of the vaso constriction of the skin.

Appendix

On the proof of Theorem 1.

(i) The function $d_j(\theta) \geq 0$ attains its minimum $d_j(\theta) = 0$ at $\theta = \hat{\theta}_j$.

Now

$$\mu_j = b'(\theta_j) = g^{-1}(\mathbf{x}_j^T\beta). \tag{5}$$

For fixed $\theta_1, \ldots, \theta_p$, let β be the solution of

$$\theta_j = b'^{-1}(g^{-1}(\mathbf{x}_j^T\beta)) \quad \text{and} \quad c_j = g(b'(\theta_j))$$

Then, if the jth row of $\mathbf{X}$ is $\mathbf{x}_j^T$, $j = 1, \ldots, p$,

$$\mathbf{X}\beta = \mathbf{c} \tag{6}$$

and, since rank $(\mathbf{X}) = p$, (6) has a unique solution.

Now choose $d > 0$ and put $d_j(\theta_j^{(d)}) = d$, $j = 1, \ldots, p$.

Since $d_j(\theta_j) \longrightarrow \infty$ as $\theta_j \longrightarrow \pm\infty$, such a $\theta_j^{(d)}$ always exists.

Put $c_j^{(d)} = g(b'(\theta_j^{(d)}))$, $j = 1, \ldots, p$ and define

$$\mathbf{x}_j^T\beta_d = c_j^{(d)},\ j = 1, \ldots, p. \tag{7}$$

Then β_d is the unique solution of $\mathbf{X}\beta = \mathbf{c}^{(d)}$.

Now let

$$c_{p+1}^{(d)} = \mathbf{x}_{p+1}^T\beta_d \quad \text{and} \quad \theta_{p+1}^{(d)} = b'^{-1}(g^{-1}(c_{p+1}^{(d)})).$$

Then either

$$d_{p+1}(\theta_{p+1}^{(d)}) < d \quad \text{or} \quad d_{p+1}(\theta_{p+1}^{(d)}) > d \quad \text{or} \quad d_{p+1}(\theta_{p+1}^{(d)}) = d. \tag{8}$$

Case I: Suppose $d_{p+1}(\theta_{p+1}^{(d)}) < d$.

Then reduce d until $d_{p+1}(\theta_{p+1}^{(d)}) = d$. To see that this is possible, note that $d_{p+1}(\theta_{p+1}^{(d)})$ is continuous in d and that choosing $d = 0$ gives

$$c_j = g(y_j),\ j = 1, \ldots, p \quad \text{and} \quad d_j = 0,\ j = 1, \ldots, p.$$

However,

$$\begin{aligned} d_{p+1}(\theta_{p+1}^{(0)}) &= d_{p+1}(b'^{-1}(g^{-1}(\mathbf{x}_{p+1}^T\beta_0)) \\ &> 0 \end{aligned}$$

since d_{p+1} has a unique minimum at $\hat{\theta}_{p+1} = b'^{-1}(y_{p+1})$ and the minimum is 0.

Case II: Suppose $d_{p+1}(\theta_{p+1}^{(d)}) > d$.

This is a special case of the general situation where $d_1, \ldots, d_{p+1}$ are not all equal.

Since b' and g are strictly monotone, d_{p+1} does not attain its minimum unless $\mathbf{x}_{p+1}^T\beta = g(y_{p+1})$.

Since $d_{p+1}(\theta_{p+1}^{(d)}) > 0$, this is not the case for β_d.

β defines a hyperplane in $(p+1)$ dimensional space. The hyperplane is horizontal with probability 0. Hence there is a direction in which β can move which will be the direction of reducing d_{p+1} . Choose the steepest such direction and move β until $d_{p+1} = d_j$ for some $j \leq p$.

Now move β in the steepest direction for which $d_{p+1} = d_j$ and the common value is reduced until a third deviance d_l satisfies $d_l = d_j = d_{p+1}$.

Continue this process until $d_j = d_{p+1}$ for all but one of the d_j, $j = 1, \ldots, p$. This case then reduces to case I.

(ii) Suppose that the value of β which minimizes $max_{j=1,\ldots,p+1} d_j(\theta_j)$ is such that $d_{p+1}(\theta_{p+1}) > d_j(\theta_j)$, $j = 1, \ldots, p$. Then, applying the procedure in (II) will always reduce the maximum deviance.