Iteratively Re-weighted Least Squares Method for Outlier Detection in Linear Regression

Nedret Billor
Cukurova University, Department of Mathematics
Faculty of Arts and Sciences
01330, Adana, Turkey
nbillor@mail.cu.edu.tr

Samprit Chatterjeer
New York University, Department of Statistics and Operations Research
44 West 4th. Street
New York, NY, 10012-1126, USA
schatter@stern.nyu.edu

Ali S. Hadi
The American University in Cairo , Department of Mathematics
P. O. Box 2511, Cairo
Cairo 11511, Egypt
ahadi@aucegypt.edu

1. Introduction

The presence of high-leverage points, individually or in groups, makes it very difficult to identify the outliers in the regression data and to obtain a robust fit. An iteratively reweighted least squares procedure as a robust fit for the standard linear model, $\mathbf{y} = \mathbf{X}\beta + \epsilon$, was proposed by Chatterjee and Mächler (1997). This method is not very effective when there is extensive masking because it is based on the standard measure of leverage p_{ii} which is affected by the masking problem. In Section **??** we present a procedure which works in the presence of masking.

2. The Method Proposed

The proposed procedure is described as follows:

Step 0: Obtain the initial weights as follows:

1. Let $(d_1, \ldots, d_n)$ be the normalized BACON distances obtained in the final step of the BACON Algorithm (Billor, Hadi, and Velleman, 2000) when it is applied to $\mathbf{X}$.

2. Let m_d be the median of $(d_1, \ldots, d_n)$. Replace d_i by $d_i = 1/max(d_i, m_d)$.

3. Compute the squared normalized version of the new d_i by

$$d_i = \frac{d_i^2}{\sum_{i=1}^{n} d_i^2}. \tag{1}$$

4. Let $\hat{\beta}^0$ be the weighted least squares estimates of the regression coefficients when using d_i in (**??**) as a weight for the ith observation.

Step j: For $j = 1, 2, \ldots$, until convergence, let $\mathbf{e}^{j-1} = \mathbf{y} - \hat{\mathbf{y}}^{j-1} = \mathbf{y} - \mathbf{X}\hat{\beta}^{j-1}$ be the residuals of the last fit. Replace e_i^{j-1} by its squared normalized version, which is given by

$$e_i^{j-1} = \frac{(e_i^{j-1})^2}{\sum_{i=1}^{n} (e_i^{j-1})^2}. \tag{2}$$

Compute $a_i = \dfrac{1 - d_i}{max(e_i^{j-1}, m_e^{j-1})}$, where m_e^{j-1} is the median of $(e_1^{j-1}, \ldots, e_n^{j-1})$ in (**??**). Finally, compute the new weights,

$$w_i^j = \frac{a_i^2}{\sum_{i=1}^{n} a_i^2}. \tag{3}$$

Let $\hat{\beta}^j$ be the weighted least squares estimates of the regression coefficients when using w_i^j as a weight for the ith observation.

As described in the full paper, the method is complemented by a simple diagnostic plot which displays clearly the nature of all the data points, distinguishing among outliers, leverage points, and well-fitted points. The proposed procedure is also illustrated by data sets which are known to have severe masking and swamping.

REFERENCES

Billor, N., Hadi, A. S., and Velleman, P. F. (2000), "BACON: Blocked Adaptive Computationally-Efficient Outlier Nominators," *Computational Statistics and Data Analysis*, 34, 279–298.

Chatterjee, S. and Mächler, M. (1997), "Robust Regression: A Weighted Least Squares Approach" *Communications in Statistics, Theory and Methods*, 26, 1381–1394.

RESUME

A new method for the detection of outliers in linear regression is proposed. The method is based on iteratively reweighted least squares. The weights depend on robust measures of residuals and leverages. This makes the method effective in dealing with the distorting effect of masking and swamping.

Nous proposons une nouvelle mthode pour la dtection des outliers en rgression linaire, base sur une itration de la mthode des moindres carrs. Les coefficients sont obtenus partir de mesures robustes des rsidus et leverages. Ceci rend la mthode efficace pour traiter les distortions dues au masking et au swamping.