A Nonparametric Robust Method for the Detection of Outliers in Linear Models

Susana Faria\(^1\), and Giuseppe Melfi\(^2\)

\(^1\) University of Minho, DMCT, 4800-058 Guimaraes, Portugal
\(^2\) Université de Neuchâtel, Groupe de Statistique, Espace de l’Europe 4, CH–2002 Neuchâtel, Switzerland

**Keywords:** Lad Regression, Robustness, Outliers, Masking Effect

1 Abstract

The detection of outliers for the standard least squares regression model is a problem which has been extensively studied. LAD regression diagnostics offers alternative approaches whose main feature is the robustness. The robustness of LAD regression to outliers are very familiar in the literature (Dodge (1987), Dodge (1997)).

In this work, we propose a nonparametric method for detecting outliers in LAD regression models and compare it to other classical methods (Hadi (1992), Rousseeuw (1987)).

We consider the standard linear model:

\[ y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad \text{for } i = 1, \ldots, n, \]

(1)

where \( n \) is the number of observations, \( x_{ij} \) is the \( j \)th explanatory variable, \( y_i \) is the response variable, \( \beta = (\beta_0, \beta_1, \ldots, \beta_p)^T \) is the column vector of unknown parameters and \( \varepsilon_i \) is the error for the observation \( i \).

The LAD regression model is determined by minimizing the sum of the absolute deviations,

\[ \sum_{i=1}^{n} \left| y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right|. \]

(2)

When a linear LAD regression model is fitted, the hyperplane always passes through at least \( p+1 \) points (Arthanari (1993)), although the solution may be non-unique. For simplicity, we assume that for every data set we deal with, the hyperplane which fits the linear LAD regression model is unique. This assumption is reasonable, in the sense that the uniqueness of the solution is very probable when the size of data sets increases. We suppose also that the data set is such that every \( p+2 \) points are not in the same hyperplane. With these assumptions the linear LAD regression model is unique and it passes through exactly \( p+1 \) points. Furthermore, if \( n > p+1 \), there is always a point which does not belong to the regression hyperplane, so having a positive absolute deviation.

Let \( S \subset \mathbb{R}^{p+1} \) be a finite discrete set of points and denote the elements of \( S \) as \((x_{i1}, \ldots, x_{ip}, y_i)\).

Consider the \( n \) data sets composed by all possible subsets of \( S \) of size \( n-1 \). Under the above assumptions for each data set we have a unique solution. For each subset we consider the LAD regression line and we give the score \( 1/m \) to the \( m \) points which maximize the absolute distance from the LAD regression line. Usually there will be only one point which maximize the absolute deviation, and this point will get a whole additional score. We define the final score of each point as the sum (over all \( n \) possible subsets of \( S \)) of scores arising from the LAD regression lines for subsets of \( S \) of size \( n-1 \). This score is produced by the repeated use of the same points, each time considering a different subset of the data set, so in a certain sense by bootstrapping the linear LAD regression model.
Based on this, we introduce a procedure for the detection of outliers in LAD regression models that appears to be not affected by masking problems.

Our procedure is illustrated and compared to others existing methods, using several data sets known to contain multiple outliers (Rousseeuw (1987), Hawkins et al. (1984)). Also, the performance of our method is investigated by a Monte Carlo study.

The data sets and the Monte Carlo indicated that the proposed procedure is effective in the detection of outliers in linear models.

It is important to note that the computation of the scores require the determination of a certain number of LAD regression models and this is computationally long. However the principle is very simple and, nowadays, the performances of common notebooks are largely sufficient to do in a fraction of seconds those computations.

References


