Simulation Modelling Practice and Theory 47 (2014) 110–120

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Least trimmed euclidean deviations for robust leverage in regression estimates

C. Chatzinakos *, G. Zioutas

Aristotle University Thessaloniki, Faculty of Engineering, Greece

ARTICLE INFO

Article history:
Received 8 April 2014
Received in revised form 3 June 2014
Accepted 5 June 2014

Keywords:
Robust regression
Outlier detection
LTED optimization
Leverage-points
Monte-Carlo simulation

ABSTRACT

Usually, in the regression models, the data are contaminated with unusually observations (outliers). For that reason the last 30 years have developed robust regression estimators. Among them some of the most famous are Least Trimmed Squares (LTS), MM, Penalized Trimmed Square (PTS) and others. Most of these methods, especially PTS, are based on initial leverage, concerning x outlying observations, of the data sample. However, often, multiple x-outliers pull the distance towards their value, causing leverage bias, and this is the masking problem.

In this work we develop a new algorithm for robust leverage estimate based on Least Trimmed Euclidean Deviations (LTED). Extensive computational, Monte-Carlo simulations, with varying types of outliers and degrees of contamination, indicate that the LTED procedure identifies successfully the multiple outliers, and the resulting robust leverage improves significantly the PTS performance.

1. Introduction

Outliers are observations that do not follow the pattern of the majority of the data. Outliers in a multivariate point cloud can be hard to detect, especially when the dimension p exceeds 2, because we can no longer rely on visual perception.

Different approaches have been proposed to over come this difficulty. Some of them are based on the minimization of a robust scale of Mahalanobis distances, the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimates Rousseeuw [14,15]. The MVE method looks for the ellipsoid with smallest volume that covers by data points where $h/2 < h < n$ for a breakdown value $(n - h)/n$. The MCD method is a highly robust estimator and its objective is to find h observations whose covariance matrix has the lowest determinant. MVE and MCD demonstrates good performance on data sets with low dimension but on larger data sets are quite computationally intensive – the exact solution requires a combinatorial search which increases rapidly with dimension. Other distance-based algorithm is the OGK estimator proposed by Maronna and Zamar [9]. However, for large dimensions the OGK is still quite computationally intensive.

In this article we present a new algorithm for detecting outliers, the new approach is completed in two stages. In the first stage we propose a new median estimate called Least Trimmed Euclidean Deviation (LTED), which is the median center of the those h points of X for which the Euclidean distance from the center is minimal.

In the second stage we use the idea of the MCD. Given the clean coverage (h points) subset of the first stage we perform only one time the concentration c-step of the MCD procedure, in order to obtain a new h-subset separated from the original X_n by an ellipsoid.

* Corresponding author. Tel.: +30 2331025985.
E-mail addresses: Chatzin@auth.gr (C. Chatzinakos), zioutas@eng.auth.gr (G. Zioutas).

http://dx.doi.org/10.1016/j.simpat.2014.06.002
1569-190X/© 2014 Published by Elsevier B.V.

The final multivariate location and scatter estimates from the ellipsoid data of the new approach, LTED, are shown to be improved comparing of the MCD estimates or competitive of others reviewed before, while requiring reasonable time. The LTED procedure are shown to perform well even under very high collinearity and contamination. We give numerical evidence indicating that we earn robustness and efficiency.

The structure of the paper is as follows: In Section 2 we describe the regression model, the outlier problem and review two of the most famous robust estimators LTS, PTS.

In Section 3 is described the penalty function and the basic tool of the PTS estimator, which is based on robust leverage. In the same section it is developed the new robust procedure LTED for leverage estimate.

Finally in Section 4, we evaluate the proposed LTED by comparing the PTS performance on simulated data, and we summarize the results in Section 5.

2. Regression

In multilinear regression models experiments data often contains outliers and bad influential observation, due to errors. It is important to identify these observations and eliminate them from the data set, since the can lead the regression estimate to take erroneous values.

Consider the linear model

$$y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + u_i \quad i = 1, \ldots, n \tag{1}$$

where data points are of the form $(\mathbf{x_i}, y_i) = (x_{i1}, \ldots, x_{ip}, y_i)$, with $x_{ip} = 1$ for regression with an intercept term. Many estimators of the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ break down in the presence of outliers. There are several kinds of outliers, for which we will follow the terminology of Rousseeuw and Zomeren [19]. A point $(\mathbf{x_i}, y_i)$ which does not follow the linear pattern of the majority of the data but whose $\mathbf{x_i}$ is not outlying is called a *vertical outlier*. A point $(\mathbf{x_i}, y_i)$ whose $(\mathbf{x_i}, y_i)$ is outlying is called a *leverage point*. We say that it is a *good* leverage point when $(\mathbf{x_i}, y_i)$ follows the pattern of the majority, and a *bad* leverage point otherwise. Summarizing, a data set can contain four types of points: regular observations, vertical outliers, good leverage points, and bad leverage points. Of course, most data sets do not have all four types.

For simple regression the data are bivariate and can be displayed in a scatterplot, so we can easily detect outlying observations by visual inspection. But for data with several explanatory variables, this is no longer possible. Then the residuals based on a robust regression detect points $(\mathbf{x_i}, y_i)$ that deviate from the linear pattern. Moreover, leverage points can be detected by computing the robust distances of $\mathbf{x_1}, \ldots, \mathbf{x_n}$ as proposed in Chatzinakos et al. [3], as will illustrated below.

The approaches to outlier identification can be separated into two categories: direct approaches and indirect approaches using residuals from the robust fit. Among famous direct approaches, Hadi and Simonoff [5] presented a procedure where it is attempted to separate the data into a set of *clean* data points and a set of points that contains the potential outliers. Atkinson [1] proposed an identification method of multiple outliers by using a simple forward search starting from initial random subsets. The procedure requires again that at least one of the subsets does not contain high-leverage outliers. Peña and Yohai [12] proposed a successful fast procedure for detecting group of outliers in many situation, where due to masking effects the usual diagnostic procedure fail. Generally, the key to the success of the above procedures is obtain a clean initial subsets of data.

An indirect approach to outlier identification is through a robust regression estimate. A famous estimator that preserves high breakdown point (HBP) is the Least Trimmed Squares (LTS) estimator of Rousseeuw and Leroy [17], that minimize the sum of the h, (coverage $h \geqslant [(n+p+1)/2]$ smallest squared residuals. But is well known that LTS loses efficiency. Some better proposals obtain high breakdown points and improve the efficiency of the LTS estimator. Among them are the S estimators of Rousseeuw and Yohai [18], the MM estimators of Yohai [24], Simpson et al. [22] and Coakley and Hettmansperger [4]. These estimators, uses a less efficient high-breakdown method as an initial estimate, and then uses an M estimation strategy based on the redescending ψ function. Although they have achieved good asymptotic properties, may have low finite-sample efficiencies if the design contains high leverage points. Morgenthaler [10] and Stefanski [23] argue that no estimator with breakdown point greater than $1/n$, can have high finite-sample efficiency in the presence of extreme leverage points. The above estimators for LTS are based mainly on the initial LTS regression of the initial coefficient estimates. Sometimes, in data contaminated by high-leverage outliers, a bad initial coefficient value does not lead to a good final robust estimation. The LTS method requires the coverage h or equivalently the number $(n-h)$ of the most likely outliers that produces the largest reduction in the residual sum of square when deleted. Unfortunately, this knowledge of h is typically unknown.

A different approach is the Penalized Trimmed Squares (PTS) proposed by Zioutas et al. [25], which does not require presenting the number $(n-h)$ of outliers to delete from the data set. The estimator PTS is defined by minimizing a convex objective function, which is the sum of squared residuals and penalty costs for discarding bad observations. The robust estimate is obtained by the unique optimum solution of the convex mathematical formula called QMIP.

The PTS estimator is very sensitive to the penalties defined a priori. In fact, these penalty costs are a function of the robust scale σ and leverage of the design points provided by LTS and minimum covariance determinant MCD of Rousseeuw and Driessen [16]. These penalties in the loss function regulate the robustness and the efficiency of the estimator.

However, due to the high computational complexity of the resulting QMIP problem, exact solutions for moderately large regression problems is infeasible. An approximate algorithm called Fast-PTS proposed by Pitsoulis and Zioutas [13] in order to compute the PTS estimator for large data sets efficiently.