Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 717–734
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

An algorithm for computing exact least-trimmed squares estimate of simple linear regression with constraints

Lei M. Li*

Department of Computational Biology and Mathematics, University of Southern California, CA 90089-1113, USA, 1042 West 36th Place, DRB 289, Los Angeles, CA 900891113, USA

Received 19 November 2003; received in revised form 2 April 2004; accepted 3 April 2004

Abstract

The least-trimmed squares estimation (LTS) is a robust solution for regression problems. On the one hand, it can achieve any given breakdown value by setting a proper trimming fraction. On the other hand, it has $\sqrt{n}$-consistency and asymptotic normality under some conditions. In addition, the LTS estimator is regression, scale, and affine equivariant. In practical regression problems, we often need to impose constraints on slopes. In this paper, we describe a stable algorithm to compute the exact LTS solution for simple linear regression with constraints on the slope parameter. Without constraints, the overall complexity of the algorithm is $O(n^2 \log n)$ in time and $O(n^2)$ in storage. According to our numerical tests, constraints can reduce computing load substantially. In order to achieve stability, we design the algorithm in such a way that we can take advantage of well-developed sorting algorithms and softwares. We illustrate the algorithm by some examples.

Keywords: Least-trimmed squares; Simple regression; Robust; Breakdown value; Constraint

1. Introduction

Statisticians routinely apply regression analysis to fit models to observations. To deal with outliers, we seek for robust and resistant regression procedures. Quite some number of perspectives exist in the literature regarding the definition of robustness.

* Tel.: +1-213-7402407; fax: +1-213-7402437.
E-mail address: lilei@hto.usc.edu (L.M. Li).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.04.003

For example, Huber (1964) studied robustness from the point of view of minimax variance. Hampel (1971, 1974) proposed the idea of influence function as an asymptotic tool to study robustness. Breakdown point is another important notion in robust analysis. Donoho and Huber (1983) defined a finite-sample version of breakdown point. Consider the classical linear model

$$y = X\beta + \varepsilon, \tag{1}$$

where $y=(y_1,\ldots,y_n)'$, $\beta=(\beta_1,\ldots,\beta_p)'$, $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_n)'$, and $X=(x_{ij})_{i=1,\ldots,n,\,j=1,\ldots,p}$. For a set of parameter β_0, we define the residuals by $r(\beta_0)=y-X\beta_0$. The least-squares estimator minimizes the sum of squares $\sum_{i=1}^{n} r_i^2(\beta)$ over β. The breakdown value of least-squares estimator is $1/n \to 0$ as $n \to \infty$. On the other hand, the highest possible breakdown point is 50%. One solution that reaches this bound of breakdown point is the least median of squares (LMS) estimator, cf. Rousseeuw (1984), which minimizes the median of squared residuals: $\min_\beta [\mathrm{med}_i\, r^2(\beta)_i]$. Unfortunately, the asymptotic efficiency of LMS is unsatisfactory because its convergence rate is only of the order $n^{-1/3}$. Another robust solution is the least-trimmed squares (LTS) estimator, which takes as its objective function the sum of smallest squared residuals; see Rousseeuw (1984). We denote the squared residuals in the ascending order by $|r^2(\beta)|_{(1)} \leqslant |r^2(\beta)|_{(2)} \leqslant \cdots \leqslant |r^2(\beta)|_{(n)}$. Then the LTS estimate of coverage h is obtained by

$$\min_{\beta} \sum_{i=1}^{h} |r^2(\beta)|_{(i)}.$$

This definition implies that observations with the largest residuals will not affect the estimate. The LTS estimator is regression, scale, and affine equivariant; see Rousseeuw and Leroy (1987, Lemma 3, Chapter 3). In terms of robustness, we can roughly achieve a breakdown point of ρ by setting $h=[n(1-\rho)]+1$. In terms of efficiency, $\sqrt{n}$-consistency and asymptotic normality similar to M-estimator exist for LTS under some conditions; see Víšek (1996, 2000) for example. Despite its good properties, the computation of LTS remains a problem.

The problem of computing the LTS estimate of β is equivalent to searching for the size-h subset(s) whose least-squares solution achieves the minimum of trimmed squares. The total number of size-h subsets in a sample of size n is $\binom{n}{h}$. A full search through all size-h subsets is impossible unless the sample size is small. Several ideas have been proposed to compute approximate solutions. First, instead of exhaustive search we can randomly sample size-h subsets. Second, Rousseeuw and Van Driessen (1999) proposed a so-called C-step technique (C stands for "concentration"). That is, having selected a size-h subset, we apply the least-squares estimator to them. Next, for the estimated regression coefficients, we evaluate residuals for all observations. Then a new size-h subset with the smallest squared residuals is selected. This step can be iterated starting from any subset. In the case of estimating a location parameter, Rousseeuw and Leroy (1987, pp. 171–172), described a procedure to compute the exact LTS solution. Rousseeuw and Van Driessen (1999) applied this idea to adjust the intercept in the