Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER Applied Mathematics and Computation 175 (2006) 609–617
APPLIED MATHEMATICS AND COMPUTATION
www.elsevier.com/locate/amc

Computational algorithm of least absolute deviation method for determining number of outliers under normality

Jong-Wuu Wu [a,*], Wen-Chuan Lee [b]

[a] *Department of Applied Mathematics, National Chiayi University, Chiayi City 60004, Taiwan, ROC*
[b] *Department of International Business, Chang Jung Christian University, Tainan, Taiwan, ROC*

Abstract

In this paper, we suggest a least absolute deviation (LAD) method for the determination of the number of upper or lower outliers in normal sample by minimizing its sample mean absolute deviation. In consequence of computation of the criterion for LAD method need to use iterative numerical analysis, we also give a computational algorithm of LAD method to find the number of upper or lower outliers in normal sample. Finally, we give two examples to study the practical performance of the LAD method.

Keywords: Least absolute deviation; Outlier; Normal distribution; Computational algorithm; Order statistic

* Corresponding author.
E-mail address: jwwu@mail.ncyu.edu.tw (J.-W. Wu).

0096-3003/$ - see front matter
doi:10.1016/j.amc.2005.07.063

1. Introduction

In the last few years, the problem of analysing data involving outliers or spurious observations has received considerable attention. Recently, Barnett and Lewis [1] propose a consecutive test to determine the number of upper or lower outliers. It is a simplified method of repeatedly applying a single-outlier procedure. In general, there are two kinds of consecutive procedure. One is called the "inward" method and is performed as follows: first, the most extreme outlier is tested; then the second most extreme; etc. The process continues until a non-significant result is obtained. The other method is the "outward" method and is performed in the reverse direction (i.e. the least extreme first). The inward method is not recommended, because it suffers seriously from masking (see Barnett and Lewis [1, Chapter 4]). Because of this, the proposed outward method can improve the situation to some extent. However, being straight-forward or simplified, these consecutive methods do not always perform well (also see examples of Section 3). In addition, Zhang and Wang [7] also used the maximum likelihood ratio test statistic T_k to determine the number of upper or lower outliers by minimizing the significance probability of the test. Next, Wu [6] proposed a least square procedure for the determination of the number of upper or lower outliers in normal sample by minimizing sample mean squares error. And he also correct of the upper bound of the significance probability of T_k suggested by Zhang and Wang [7]. Suppose $X_1, X_2, \ldots, X_n$ are random sample which $n-k$ of them follow a normal distribution and the remaining k are outliers with reference to the hypothesised distribution. Here, both the parameters of the distribution and k are unknown. Let $X_{1:n} \leqslant X_{2:n} \leqslant \cdots \leqslant X_{n:n}$ denote the order statistics corresponding to the random sample $X_1, X_2, \ldots, X_n$. Suppose that the underlying distribution is a normal distribution $N(\mu, \sigma^2)$ with cumulative distribution function (c.d.f.) as

$$F(x;\mu,\sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad -\infty < x, \ \ \mu < \infty, \ \ \sigma > 0, \tag{1}$$

where the parameters μ and σ are unknown and Φ is c.d.f. of the standard normal distribution.

Suppose that $X_{n-k+1:n}, X_{n-k+2:n}, \ldots, X_{n:n}$ are k upper outliers and that they belong to the normal distribution $N(\mu+\lambda\sigma, \sigma^2)$, $\lambda > 0$. Here, we permit k in the interval $[0, n - \left[\frac{n}{2}\right] - 1]$.

In this paper, we propose a least absolute deviation (LAD) method to determine the number of upper outliers in a sample of size n from the normal distribution by minimizing sample mean squares error. In consequence of computation of the criterion for LAD method need to use iterative numerical analysis, we also give a computational algorithm of LAD method to find the number of upper outliers in normal sample. Moreover, this method is free from the effects of masking and swamping, when testing upper outliers in normal