Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 3955–3967
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Bayesian identification of clustered outliers in multiple regression

Donna L. Mohr*

Department of Mathematics and Statistics, University of North Florida, Jacksonville, FL 32224, USA

Received 6 October 2004; received in revised form 25 July 2005; accepted 6 April 2006
Available online 2 May 2006

Abstract

We propose a Bayesian model for clustered outliers in multiple regression. In the literature, outliers are frequently modeled as coming from a subgroup where the variance of the errors is much larger than in the rest of the data. By contrast, when a cluster of outliers exists, we show that it can be more informative to model them as coming from a subgroup where different regression coefficients hold. We can explicitly model the clustering phenomenon by assuming that the probability of an outlier is a function of the explanatory variables. Fitting proceeds via the Gibbs sampler, using the Metropolis–Hastings algorithm to produce variates from the more unusual distributions. Initialization uses a least median of squares fit, and in some ways this method can be viewed as a Bayesian version of the many algorithms that use this fit as a start to some more efficient estimator. This method works very well in a variety of test data sets. We illustrate its use in a data set of sailboat prices, where it yields information both on the identity of the outliers and on their location, spread, and the regression coefficients inside the minority subgroup.

Keywords: Switching regression; Least median of squares; Gibbs samplers; Metropolis–Hastings algorithm

1. Introduction

The difficulties that clusters of outliers cause in regression have been well documented, and a number of alternatives to least squares regression have been proposed as a solution. Rousseeuw's (1984, referred to herein as R-84) least median of squares (LMS) regression is an example of a high-breakdown estimator designed to resist clusters of influential outliers. However, as discussed below, even this method can give some surprising answers when numerous outliers lie in a compact cluster.

To understand the strengths and weaknesses of LMS, it is instructive to compare an example presented in R-84 with the slight revision of it presented by Justel and Peña (2001, referred to as JP-2001). In the 50 observations of R-84, the 30 inliers come from a model where $y = 2 + x + \varepsilon$, where the ε are normally distributed with a mean of 0 and a standard deviation of 0.2. The 20 outliers come from a spherical bivariate normal distribution centered at mean (7, 2), each component having a standard deviation of 0.5. A sample realization from this process is shown in Fig. 1, together with the LMS fitted line. The JP-2001 example is very similar, but now the outliers also have a standard deviation of 0.2, forming a more compact cluster than in the original data. Their data is shown in Fig. 2, together with the LMS fitted line. The decrease in dispersion in the outlier cluster had a tremendous impact on the LMS fit. Since the cluster is so compact and such a large proportion of the data set (40%), the LMS line is able to find a few points over in the

* Tel.: +1 904 6202884; fax: +1 904 6202818.
E-mail address: dmohr@unf.edu.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.04.008

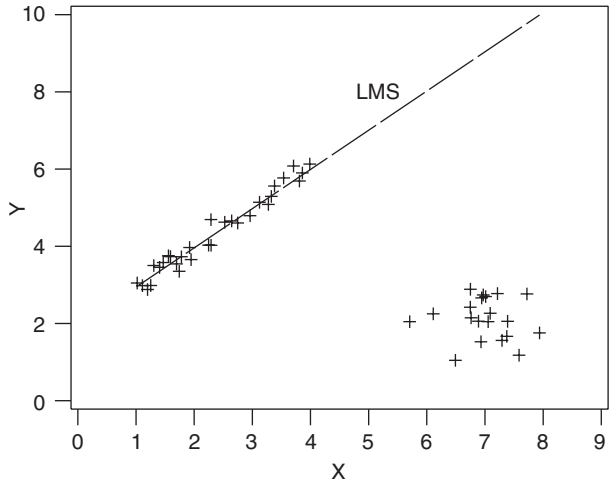

Fig. 1. R-84 example. The LMS line successfully passes through the Group 0 (inlier) set.

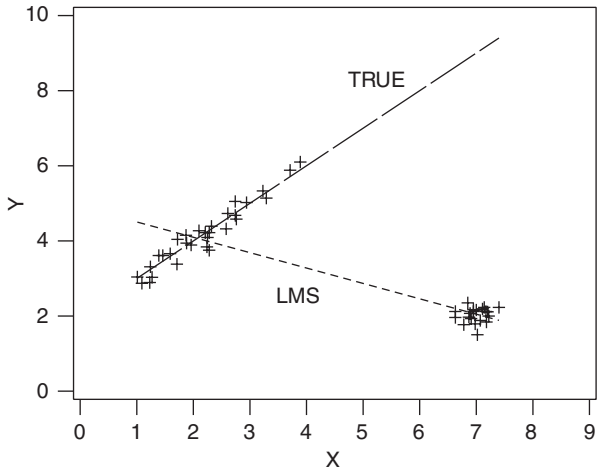

Fig. 2. JP-2001 example. The LMS fit reflects the compact cluster of outliers.

inlier group which, when put together with the outlying cluster, allow it to form a new line with very good fit through 50% of the data.

Justel and Peña (1996) proposed a Bayesian scale contamination (or variance-inflation) model for analyzing regression outliers. In such a model, the outliers are assumed to come from a normal distribution with the same mean as the inliers, but a higher variance. Their initial work showed that when large influential clusters of outliers are present (as in R-84), straightforward implementation of Bayesian mixture models using Gibbs sampling suffers from the same problems of masking and swamping that plague least-squares and M-estimators. Returning to this problem, JP-2001 proposed a solution based on careful initialization with a good candidate set of outliers (identified through a pilot run of numerous short Gibbs sampler chains). Their analysis of the JP-2001 data correctly separated the inliers and outliers.

However, the variance-inflation mechanism is better suited to generating scattered outliers rather than clusters, and this may explain why obtaining the correct solution is fairly delicate. To illustrate, Fig. 3 shows the number of correctly classified observations (black line) if we initialize a sampler for the JP model at the Potential Outlier set they recommend from Stage 2, and continue through 30,000 iterations. Initially, because of the good starting values, almost all observations are correctly classified. Sooner or later (depending on the sequence of random numbers) the chain will suddenly shift to a state where most true outliers are masked, and many inliers appear as outliers. The same chart shows