COMPUTATIONAL STATISTICS & DATA ANALYSIS

ELSEVIER Computational Statistics & Data Analysis 36 (2001) 351–382

www.elsevier.com/locate/csda

A Comparative analysis of multiple outlier detection procedures in the linear regression model

James W. Wisnowski[a,*], Douglas C. Montgomery[b], James R. Simpson[c]

[a]*Department of Mathematical Sciences, US Air Force Academy, 2354 Fairchild Dr Suite 6D2A, USAF Academy, CO 80840-6252, USA*

[b]*Department of Industrial and Management Systems Engineering, Arizona State University, Tempe, AZ 85287-5906, USA*

[c]*Department of Industrial Engineering, Florida A&M University-Florida State University, Tallahassee, FL 32310-6046, USA*

Received 1 January 1999; received in revised form 1 August 2000

Abstract

We evaluate several published techniques to detect multiple outliers in linear regression using an extensive Monte Carlo simulation. These procedures include both direct methods from algorithms and indirect methods from robust regression estimators. We evaluate the impact of outlier density and geometry, regressor variable dimension, and outlying distance in both leverage and residual on detection capability and false alarm (swamping) probability. The simulation scenarios focus on outlier configurations likely to be encountered in practice and use a designed experiment approach. The results for each scenario provide insight and limitations to performance for each technique. Finally, we summarize each procedure's performance and make recommendations. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Outlier; Multiple outliers; Robust regression; Minimum volume ellipsoid; Monte Carlo simulation

1. Introduction

There has been considerable interest in recent years in the detection and accommodation of multiple outliers in statistical modeling. This paper uses Monte

[*] Corresponding author.

E-mail address: jim.wisnowski@usafa.af.mil (J.W. Wisnowski).

PII: S 0167-9473(00)00042-6

Carlo simulation to evaluate numerous recently published outlier techniques in the linear regression model. Kianifard and Swallow (1990) report a similar smaller study using a few earlier techniques. Other comparative analyses typically appear in journal articles where authors propose a new methodology; however, these studies are often limited in scope and breadth of techniques. Our approach tests the latest and most respected representative multiple outlier detection procedures across a number of realistic and challenging regression scenarios.

In general, Barnett and Lewis (1994) define outliers as observations that appear inconsistent with the remainder of the data set. For this paper, we wish to identify outliers in *linear regression* modeling. Specifically, we are concerned with observations that differ from the regression plane defined by the bulk of the data. It is important to identify these types of outliers in regression modeling because the observations, when undetected, can lead to erroneous parameter estimates and inferences from the model. Additionally, these outliers may be of interest themselves to provide insight into process behavior at certain operating conditions.

If there is only a single or a few outliers, many standard least-squares regression diagnostic quantities and plots will reliably identify these observations. These diagnostics have been shown to fail in the presence of multiple outliers, particularly if the observations are clustered in an outlying cloud. The measures may either fail to identify the outliers (masking), identify the clean observations as outliers (swamping), or could both mask and swamp observations. To overcome the limitations of the standard least-squares diagnostics, numerous multiple outlier detection techniques have been proposed to identify the outlying subset of observations.

The outlying observations can be remote in the levels of the regressor or explanatory variables (exterior X-space observations). These are considered high-leverage points because they are influential and pull the regression plane toward them. We refer to cases that are not unusual in X-space as interior X-space observations. Further classification of outliers is possible with respect to the regression model. If the observations do not follow the regression surface from the bulk of the data, then these cases are known as regression outliers. We are concerned with two main outlier configurations likely to be encountered in practice: (1) observations that are interior X-space regression outliers and (2) observations that are exterior X-space regression outliers. It turns out that some multiple outlier detection procedures work well because the outliers' response values are distant from the range of the clean cases' responses. An example data set would be where the clean response values range between 50 to 100 and the outliers' response values range between 800 to 1000. We consider testing in scenarios that have interior and exterior regression outliers where the outliers' response values are unusual and also in scenarios where they are not unusual with respect to the clean response values. A third important outlier scenario occurs when the observations are remote in X-space but the response values are in-line with the regression surface. We limit the scope of this paper by not including these high-leverage "good outliers".

Section 2 briefly describes the multiple outlier detection procedures used in this comparative study. Detailed summaries of many of these and other multiple outlier detection procedures can be found in Hadi and Simonoff (1993), Barnett and Lewis