ELSEVIER Computational Statistics & Data Analysis 33 (2000) 361–380

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

On robust linear regression with incomplete data

A.C. Atkinson*, Tsung-Chi Cheng

Department of Statistics, London School of Economics, Houghton Street,
London WC2A 2AE, UK

Received 1 February 1998; received in revised form 1 April 1999

Abstract

In this paper, we use recently developed methods of very robust regression to extend missing value techniques to data with several outliers. Simulation experiments reveal that additional outliers may be imputed if one ignores the outliers already in the data. The combination of the forward search algorithm for high breakdown point estimators and the EM algorithm or multiple imputation for missing values can avoid problems of this kind. Some multiple deletion diagnostics for linear regression with incomplete data are also discussed. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: EM algorithm; Forward search algorithm; High breakdown point; Least trimmed squares; Missing values; Multiple imputation; Regression diagnostics; Stalactite plot

1. Introduction

The study of missing values is one of the most important topics in applied statistics, especially in survey problems and medical and biological data. In this paper, we use recently developed methods of very robust regression to extend missing value techniques to data with several outliers. The usual assumption is that missing values are "missing at random" (MAR) (Rubin, 1976; see also Little and Rubin, 1987): the missing-data mechanism does not depend on X_{mis} (the set of missing values) though it may possibly depend on X_{obs} (the set of observed values). If the missingness mechanism does not depend on the parameters of the model, this assumption is called distinct. Moreover, if both MAR and distinctness hold, then the missing-data

* Corresponding author. Tel.: 00-44-171-955-7622; fax: 00-44-171-955-7416.
E-mail address: a.c.atkuson@lse.ac.uk (A.C. Atkinson)

PII: S 0167-9473(99)00061-4

mechanism is said to be ignorable (Little and Rubin, 1987; Rubin, 1987). Little (1992) suggests that model-based methods, such as maximum likelihood (ML), Bayesian methods and multiple imputation (MI), are best among the current methods for dealing with missing values.

The EM algorithm (Dempster et al., 1977) is an iterative computational method to get a maximum-likelihood estimate when the data can be conveniently viewed as incomplete. It has been widely used to cope with missing data problems. However, it does not provide the variance–covariance estimate of estimated parameters for the linear regression model with incomplete data. There are several approximate methods to get the asymptotic standard errors of ML estimators, for example, scoring or Newton algorithm, bootstrapping the sample, or constructing numerical approximations to the information matrix by the EM computations (see Little, 1992). Beale and Little (1975) gave an approximate formula for estimating the covariance matrix, which has quite stable performance through different missing patterns in a simulation study (Little, 1979). Rubin (1987) proposed multiple imputation, that requires multiply-imputed values for each missing value, resulting in multiple completed data sets. One of the advantages of this method is to avoid underestimation of the true variance.

Now consider the effects of outlying cases. The E-step of the EM algorithm which involves filling in missing values is based on the expected values of the data. Under the normal model of applying multiple imputation, the distribution of missing elements is defined by the multivariate normal linear regression of the missing variables on the observed variables (see Rubin and Schafer, 1990). However both of them are affected by outlying cases. We may therefore impute extra outliers if the existing outliers are ignored. The masking and swamping phenomena are more serious in incomplete data than those without missing values.

For the detection of multiple outliers from linear regression problems without missing data, it essentially needs the high breakdown estimators, such as the least median of squares (LMS) and least trimmed squares (LTS) (see Rousseeuw, 1984; Rousseeuw and Leroy, 1987). A problem with these high breakdown estimators is the lack of efficient algorithms. Several algorithms have been proposed recently (e.g. Atkinson, 1994; Hawkins, 1994; Atkinson and Cheng, 1999). Among these newly developed methods, Atkinson's (1994) forward search algorithm for the LMS is comparatively fast. Atkinson and Cheng (1999) adapt the forward search algorithm for the LTS, which maintains a high breakdown point, to resist the contamination of data, as well as to keep a high efficiency. For the outlier problems in missing values, Shih and Weisberg (1986) extended the distance measure (Cook and Weisberg, 1980) of assessing the influence of the ith case by deleting it from the model for incomplete data. Some quantities of multiple diagnostics will be discussed in this paper. However they are less attractive and limited in the problems of high dimension and large sample size. A procedure using estimators with a high breakdown point is then proposed to detect multiple outliers for the linear regression model with incomplete data. The main idea of the algorithm follows the forward search algorithm, that starts with a randomly selected subset of observations. The observations of the subset are incremented in such a way that outliers are likely to be excluded. If the data are