Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Statistics & Data Analysis 50 (2006) 1840–1854

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Influential data cases when the C_p criterion is used for variable selection in multiple linear regression

S.J. Steel, D.W. Uys*

Department of Statistics and Actuarial Science, Stellenbosch University, Private Bag X1, 7602 Matieland, South Africa

Received 29 June 2004; received in revised form 8 December 2004; accepted 8 February 2005
Available online 14 March 2005

Abstract

The influence of data cases when the C_p criterion is used for variable selection in multiple linear regression analysis is studied in terms of the predictive power and the predictor variables included in the resulting model when variable selection is applied. In particular, the focus is on the importance of identifying and dealing with these so-called selection influential data cases before model selection and fitting are performed. A new selection influence measure based on the C_p criterion to identify selection influential data cases is developed. The success with which this influence measure identifies selection influential data cases is evaluated in two example data sets.

Keywords: C_p criterion; Influential data cases; Multiple linear regression; Variable selection

1. Introduction

Multiple linear regression analysis is a widely used and well documented statistical procedure. Two aspects of regression analysis which have been particularly well investigated are identifying and dealing with influential data cases, and selecting a subset of the explanatory variables for use in the regression function. Standard references on the first issue include Cook (1977), Belsley et al. (1980) and Atkinson and Riani (2000),

* Corresponding author. Tel.: +27 21 808 3244; fax: +27 21 808 3830.
E-mail address: dwu@sun.ac.za (D.W. Uys).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.02.003

while Burnham and Anderson (2002) and Miller (2002) provide recent overviews of the second issue.

Although influential data cases and variable selection have separately been extensively dealt with in the literature, relatively little have been published on investigations into a combination of these two problems. We briefly refer to some of the relevant references. Chatterjee and Hadi (1988) propose measuring the effect of simultaneous omission of a variable and an observation from the data set in terms of changes in the values of the least squares regression coefficients, the residual sum of squares, the fitted values, and the predicted value of the omitted observation. Peixoto and Lamotte (1989) investigate a procedure which adds a dummy variable for each observation to the explanatory variables. Variable selection is then performed, and observations corresponding to selected dummy variables are pronounced to be influential. Léger and Altman (1993) identify conditional and unconditional approaches to the problem of identifying influential data cases in a variable selection context. In the *conditional* approach the full data set is used to select a set of explanatory variables, and case diagnostics are then calculated conditional on this model, i.e., the set of selected variables remains fixed. In the *unconditional* approach we apply variable selection to the full data set and calculate a vector of fitted values; we then omit the data case under consideration from the data set and repeat the variable selection as well as calculation of the vector of fitted values; finally, a standardised distance between the two vectors of fitted values is calculated to measure the influence of the omitted case. Léger and Altman (1993) argue that the unconditional approach is preferable since it explicitly takes the variable selection into account when trying to quantify the influence of a given data case. Arguing along similar lines, Hoeting et al. (1996) point out that the model which is selected can depend upon the order in which variable selection and outlier identification are carried out. They therefore propose a Bayesian method which can be used to simultaneously select variables and identify outliers.

In this paper we restrict attention to variable selection using the C_p statistic proposed by Mallows (1973). Our contribution is the introduction of a new p-value based procedure for identifying influential data cases in this context. Weisberg (1981) shows how the C_p statistic can be written as a sum of n terms (where n is the number of data cases), with each term in the sum corresponding to one of the n cases. In Section 2 of this paper we provide a brief exposition of the coordinate free approach to linear model selection, and in Section 3 we will see that the breakup of the C_p statistic described by Weisberg (1981) can also be formulated within the coordinate free approach. Section 4 of the paper is devoted to a discussion of the p-value based procedure for identification of influential data cases in a variable selection context, and Section 5 contains two examples illustrating application of the procedure. We close in Section 6 with conclusions and open questions.

2. A coordinate free approach to linear model selection

The coordinate free approach to variable selection in multiple linear regression analysis offers the advantage that the results which are obtained can also be applied in a wider linear model context. In this section we briefly indicate how the C_p statistic can be derived within