RESEARCH INTO MULTIPLE OUTLIERS IN LINEAR REGRESSION ANALYSIS

Barış Aşıkgil*† and Aydın Erar*

Received 23:11:2006 : Accepted 08:06:2009

Abstract

Studying the observations in regression analysis it is seen that the output of regression is affected from outliers in the direction of the dependent and / or the independent variables. In this paper multiple outliers are examined in two real data sets. The results concerned with which method can determine multiple outliers better are examined with the help of some statistics and REC curve which can be used for determining efficiency. Also, the results are tried to support by using Monte Carlo Simulation.

Keywords: Regression, Multiple outliers, Forward search, Stalactite plot, REC curve.


1. Introduction

The aim of regression analysis is to form a suitable model providing an explanation of the relationship between the dependent (Y) and independent (Xj) variables with the help of data. To form this model using the ordinary least squares (OLS) method some assumptions have to be satisfied. Specifically, the errors must have zero mean and equal variance, and be uncorrelated. If an inference is made, they must be normally distributed. Also, there must be no complete or approximate multicollinearity between the independent variables. The model of linear regression is given in matrix notation as follows:

\[ Y = X \beta + \varepsilon. \]

The observation distances matrix (projection or hat matrix) is given by

\[ H = X (X'X)^{-1} X'. \]