Procedures for Detecting Multiple Outliers in Linear Regression Using R*

Soon Sun Kwon[†], Gwi Hyun Lee[‡] and Sung Hyun Park[§]

Abstract

In recent years, many people use R as a statistics system. R is frequently updated by many R project teams. We are interested in the method of multiple outlier detection and know that R is not supplied the method of multiple outlier detection. In this talk, we review these procedures for detecting multiple outliers and provide more efficient procedures combined with direct methods and indirect methods using R.

Key Words : R, Multiple Outlier Detection, Direct Method, Indirect Method

1 Introduction

R is used excellent software in recent years in the statistics, calculation and graphical display. It provides a flexible system for data analysis that can be extended as needed. And it has an effective data handling and storage facility, a well developed, simple end effective programming language which includes conditionals, loops, user defined recursive functions and input output facilities. R is a language for statistical computing similar to the S language developed at Bell Laboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. There is an important difference between R and the other main statistical systems(S, SAS, SPSS). SAS and SPSS will copious output from a regression or discriminant analysis. But, R will give minimal output and store the results in a fit object. S is similar to R, yet S is supplied as a commercial package. In this talk, we propose the detection of multiple outlier using R. We know that S supplies some detection methods, but we propose more detection methods as R.

2 Review of Multiple Outlier Procedures

The multiple outlier detecting procedures for linear regression are direct procedures and indirect procedures. The direct methods use algorithms to isolate outliers and the indirect methods use the results from robust regression estimates. Both the direct and indirect procedures considers the linear model

$$y = X\beta + \epsilon$$

*This work was supported by a grant R01-2003-000-10220-0 from the Korea Science and Engineering Foundation.

[†] Ph.D, Dept. of Statistics, Seoul National University, Seoul 151-747, Korea

[‡] M.S, Dept. of Statistics, Seoul National University, Seoul 151-747, Korea

[§] Professor, Dept. of Statistics, Seoul National University, Seoul 151-747, Korea

where y is the response vector of dimension n, X is the $n \times p$ matrix of regressor variables with intercept, ϵ is the column vector of n random errors assumed to have mean 0 and covariance matrix $\sigma^2 I$.

2.1 Direct procedures

The direct procedures are based on either sequential deletion (backward search) of outlying observations or sequential addition (forward search) of clean observations. Generally, methods using forward search outperform backward methods. We consider the forward search procedures from Hadi and Simonoff (1993) and Swallow and Kianifard (1996). And we consider the direct procedure based on the eigenstructure of the influence matrix from Pena and Yohai(1995) and the clustering algorithm from Sebert et al. (1998). The Hadi and Simonoff forward (1993)search algorithm initially determines a clean subset of $(p+1)$ observations from the smallest absolute value of the adjusted residual from a least- squares fit, $a_i = e_i/\sqrt{(1-h_{ii})}$. And then test the outlying of the remaining points relatives to the clean subset. The Swallow and Kianifard (1996) suggest recursive residuals standardized by a robust estimate of scale as a test statistic to classify multiple outliers. The Pena and Yohai (1995) influence matrix algorithm uses the eigenstructure of an influence matrix which is defined as the matrix of uncentred covariances of the effect on the whole data set of observation. The Sebert et al.(1998) clustering algorithm uses a single linkage clustering algorithm with the Euclidean distances for the standardized predicted and standardized residual values from a least-squares fit.

2.2 Indirect procedure from robust regression estimators

Robust regression techniques accommodate outliers by downweighting or ignoring the unusual observations to ensure they are not too influential on the regression parameter estimates. The multiple outlier detection of common robust regression estimators is tested in below methods. The common robust estimators are Least Median of Squares (LMS), Least Trimmed sum of Squares (LTS), and M-estimators. We also consider the MM-estimator from Yohai (1987), standard generalized M-estimator, Coakley and Hettmanamsperger (1993) and the Simpson and Montgomery estimator (1998). Rousseuw (1984)introduced the high breakdown LMS estimators. And he proposed the high breakdown LTS estimator as an efficient alternative to LMS. But, a disadvantage of the LTS method is its lack of efficiency because of very slow convergence. Huber (1973) developed the M-estimator by minimizing a symmetric function of the residuals over the parameter estimates. The MM-estimator is a high-breakdown and high-efficiency estimator with three stages. The Standard generalized M-estimator improves the usual M-estimator by taking into account the leverage as measured by hat diagonals. The Coakley and Hettmansperger estimator uses LTS as initial estimate and adjust the estimates with empirically determined weights. The Simpson and Montgomery estimator uses a high-breakdown S-estimate for the initial estimate that minimizes the dispersion of the residuals. A related approach to the indirect methods from the robust regression estimators is the Rousseeuw and van Zomeren (1990) multiple outlier detection procedure.

3 Implementation of multiple outlier procedures using R

We obtained below classic multiple outlier data sets. We use R to compare methods of multiple outlier procedures. And we propose a modified Sebert procedure which use robust residuals and predicted values instead of standardized residuals and predicted values from OLS.

Table 1: Classic multiple outlier data sets

No	Data sets	*k*	*n*	Outliers
1	Telephones Data (Rousseeuw and Leroy, 1987)	1	24	15 24
2	Hertzsprung-Russell Stars Data (Rousseeuw and Leroy, 1987)	1	47	11,20,30,34
3	Hawkins, Bradu and Kass Data (Hawkinset al., 1984)	3	75	1 14
4	Hadi, Simonoff Data (Hadi and Simonoff, 1993)	2	25	1,2,3
5	Modified Wood Gravity Data (Rousseeuw and Levoy, 1987)	5	20	4,6,8,19
6	Stackloss Data (Brownlee, 1965)	3	21	1 4,21
7	Body and Brain Weight Data (Rousseeuw et al.1990)	1	28	6,16,25

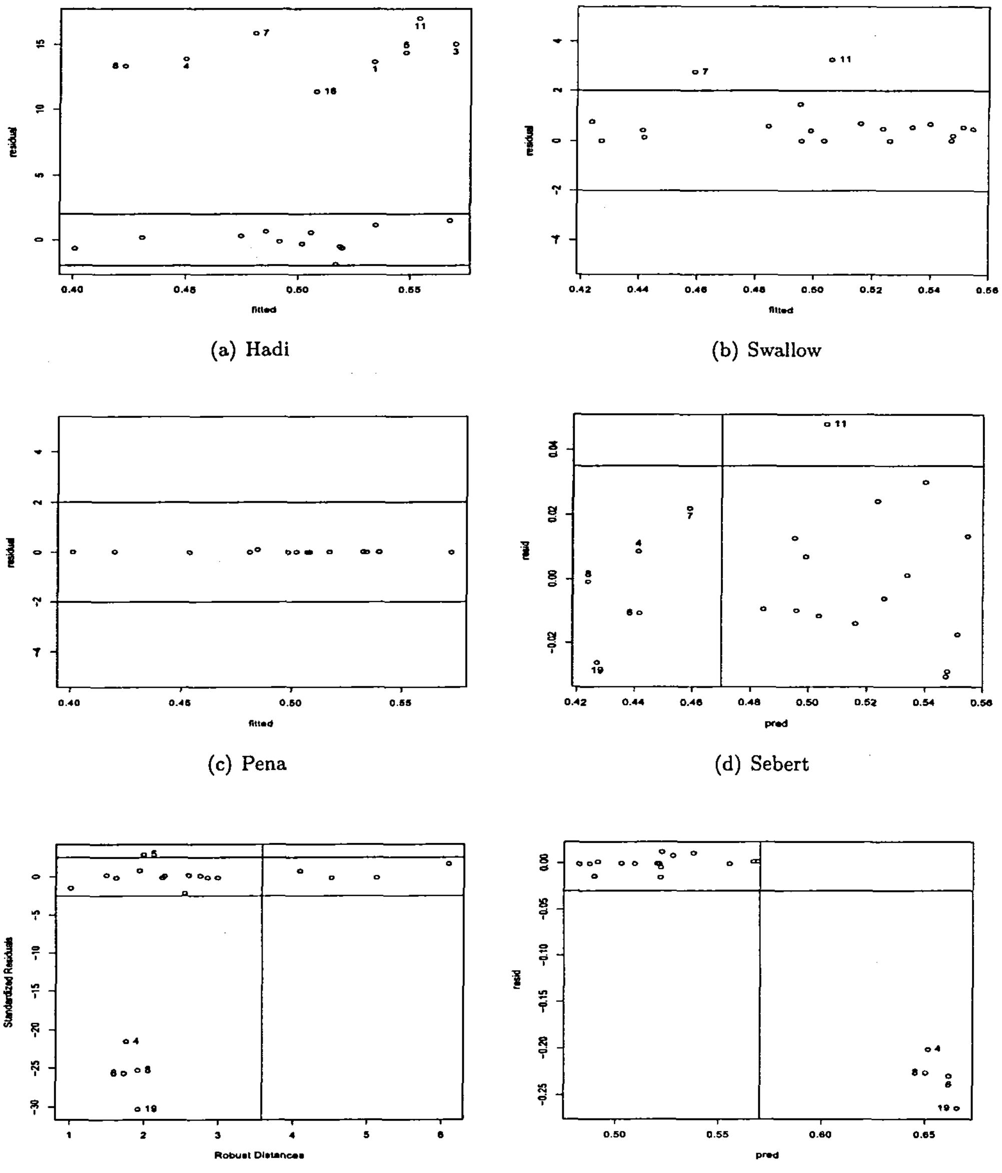
Table 2: Comparison of multiple outlier procedures

Data sets	Hadi	Swallow	Pena	Sebert	Rousseeuw	Modified Sebert
Telephones	5-11,16,21-24	15-20	.	15-24	14-21	15-20
Hert.-Roussell	1,2,4,5 6,8-13,20,30 32,33,36 38,39,40	11,20,30,34	11,20,30,34	7,11,14 20,30,34	7,9,11 20,30,34	7,11,14 20,30,34
Hawkins	1-10	1-10	1-10	1-14	1-10	1-10
Hadi	1,2,3,6 9,11-13,17 19,20,24	1	1,2,3	1,2,3,12	1-3,6,11 12,13,17,24	12
Wood Gravity	1,3,4,5 7,8,11,16	7,11	.	4,6,7 8,11,19	4,5,6,8,19	4,6,8,19
Stackloss	5,7,8,13 14,17,18,20,21	.	1,3,4	1,2,3,4,21	1-4,13,14,20,21	1-4,21
Body	2,6,9,12 14-17,24,25,28	7,15	7,15	7,15,25	2,6,7,9 12,14-16,24,25	25

We suggest detection of multiple outliers combined with graphical display using R. We display graph about Modified Wood Gravity Data.

4 Concluding Remarks

We provide the methods that either most commonly used and recently proposed the detection of outliers in linear regression data using R. In this paper, we propose the six methods of multiple outlier detection using R. But, each method has advantage and disadvantage. Thus, we study further performances of multiple outlier procedures using R.



(a) Hadi

(b) Swallow

(c) Pena

(d) Sebert

(e) Rousseeuw

(f) Modified Sebert

Figure 1: Outlier Detection graph

References

[1] Coakley and Hettmansperger(1993). A bounded influence, high breakdown, efficient regression estimator,*JASA*,**88**, pp.872-880.

[2] Hadi and Simonoff(1993). Procedures for the identification of multiple outliers in linear models, *JASA*, **88**, pp.1264-1271.

[3] Kianifard and Swallow(1990). A monte carlo comparison of five procedures for identifying outliers in linear regression, *Commun. Statist. Part A Theory methods*,**19**, pp.1913-1928.

[4] Pena and Yohai(1995). The Detection of influential subsets in linear regression by using an influence matrix, *J. R. Statist. Soc. B*, **57**, pp.145-156.

[5] Rousseeuw, P.J.(1984). Least median of squares regression.*JASA*, **79**, pp.871-881.

[6] Rousseeuw and van Zomeren(1990). Unmasking multivariate outliers and leverage points". *JASA*,**85** pp.633-639.

[7] Sebert, Montgomery and Rollier(1998). A clustering algorithm for identifying multiple outliers. *CSDA*, **27**, pp.461-484.

[8] Swallow and Kianifard(1996). Using robust scale estimates in detecting multiple outliers in linear regression. *Biometrics*,**52**, pp.545-556.

[9] Wisnowski, Montgomery and Simpson(2001). A comparitive analysis of multiple outlier detection procedures in the linear regression model. *CSDA*, pp.351-382.