The Korean Communications
in Statistics Vol. 10 No. 3, 2003
pp. 1037-1046

An Equivariant and Robust Estimator in Multivariate Regression Based on Least Trimmed Squares[1]

Kang-Mo Jung[2]

Abstract

We propose an equivariant and robust estimator in multivariate regression model based on the least trimmed squares (LTS) estimator in univariate regression. We call this estimator as multivariate least trimmed squares (MLTS) estimator. The MLTS estimator considers correlations among response variables and it can be shown that the proposed estimator has the appropriate equivariance properties defined in multivariate regression. The MLTS estimator has high breakdown point as does LTS estimator in univariate case. We develop an algorithm for MLTS estimate. Simulation are performed to compare the efficiencies of MLTS estimate with coordinatewise LTS estimate and a numerical example is given to illustrate the effectiveness of MLTS estimate in multivariate regression.

KeyWords : Breakdown point; Equivariance; Least trimmed squares estimator; Multivariate regression; Outliers.

1. Introduction

In linear regression model the least squares (LS) estimator is most well known, because it is simple and has the closed form solution to a certain systems of linear equations. Under the assumption of normality of random errors the LS estimator is optimal. However, LS estimator is very sensitive to outliers. In fact even a single outlier may destroy LS estimate. Many alternative methods in univariate linear regression models have been proposed. M, GM estimators are commonly used (Hampel et al., 1986), but the breakdown points of these estimators cannot exceed the inverse of the dimension of explanatory variables space. The least median of squares (LMS) and least trimmed squares (LTS) estimators (Rousseeuw and Leroy, 1987) have 50% breakdown point, but a low asymptotic efficiency. A vast amount of literature has treated robust estimators in univariate linear regressions. In multivariate regression model Rao (1988) used univariate least absolute deviation regression separately for

1) This work was supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF).
2) Associate professor, Department of Informatics & Statistics, Kunsan National University, Kunsan 573-701 ,Korea.
kmjung@kunsan.ac.kr

each response. Chakraborty (1999) suggested a new extension of least absolute deviation regression based on the so-called transformation and retransformation method. Also Rousseeuw et al. (2001) and Ollila et al. (2002) proposed robust multivariate regression estimators based on robust estimation of the joint location vector and scatter matrix of the explanatory and response variables.

In this paper we propose an affine and robust estimator of regression parameters in multivariate linear regression model. This estimator is based on LTS estimator in univariate regression model, which is a most popular high breakdown estimator and more efficient than LMS estimator. Even though LS estimator is not robust, it is affine equivariant under nonsingular linear transformations of the response variables. The lack of this property makes estimator practically meaningless when the values of regression variables are measured in different scales. The use of univariate regression estimator for each coordinate of the response vector, for example Rao (1988), does not take into account of correlations among response variables. It is called a coordinatewise estimator. Moreover such an approach in multivariate linear regression models does not assure the affine equivariance. The estimator proposed in Section 2 adopts transformation and retransformation approach (Chakraborty and Chaudhuri, 1998; Chakraborty, 1999) for regression equivariance and it also uses a covariance matrix of error vectors. The new estimator has 50% breakdown point, because it inherits the breakdown point of LTS estimator.

In Section 2 we define a multivariate regression model and propose a new estimator. We call this estimator as multivariate least trimmed squares (MLTS) estimator. We develop an algorithm to compute MLTS estimate. In Section 3 we describe statistical properties of the estimator such as breakdown point and affine equivariance in multivariate linear regression model. In Section 4 simulation and a numerical example are given to illustrate the effectiveness of our proposed estimate. Simulation results show that MLTS estimate appears to be more efficient than coordinatewise LTS estimate when there exist correlations among variables of error vector. From a numerical example we observe that MLTS estimate gives very useful influence information about observations in multivariate regressions.

2. Multivariate Least Trimmed Squares Estimator

Consider the multivariate linear regression model

$$\boldsymbol{y}_i = \boldsymbol{B}^T \boldsymbol{x}_i + \boldsymbol{e}_i \;; \qquad i = 1, \ldots, n, \tag{1}$$

where the size of response vector $\boldsymbol{y}_i$ is d, the length of regressor $\boldsymbol{x}_i$ is p, $\boldsymbol{B}$ is a $p \times d$ matrix of unknown coefficient parameters, and $\boldsymbol{e}_i$s are random errors uncorrelated with $\boldsymbol{x}_i$. The first element of $\boldsymbol{x}_i$ is one, so the number of regressor variables is $(p-1)$. The $\boldsymbol{e}_i$s are independently and identically distributed. Assume that $\mathrm{cov}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}$ is nonsingular.

Let α be a subset of $\{1, \ldots, n\}$ having $d+p$ elements. Write $\alpha = I \cup J$ where $I = \{i_1, \ldots, i_p\}$ and $\{j_1, \ldots, j_d\}$. Let $\boldsymbol{X}_I$ be the $p \times p$ matrix whose k-th row vector is the i_k-th row vector of the $n \times p$ data matrix $\boldsymbol{X}$. Similarly the $p \times d$ matrix $\boldsymbol{Y}_I$, the $d \times p$ matrix $\boldsymbol{X}_J$ and the $d \times d$ matrix $\boldsymbol{Y}_J$ are defined. Assume that $\boldsymbol{X}_J$ is nonsingular. Define

$$\boldsymbol{E}_\alpha^{\ T} = \boldsymbol{Y}_J - \boldsymbol{X}_J \ \boldsymbol{X}_I^{-1} \ \boldsymbol{Y}_I. \tag{2}$$

The matrix $\boldsymbol{E}_\alpha$ is assumed to be invertible, and we define the transformation response vectors as $\boldsymbol{w}_l(\alpha) = \boldsymbol{E}_\alpha^{-1} \ \boldsymbol{y}_l$ for $1 \le l \le n$ and $l \notin \alpha$. We apply univariate LTS regression on each coordinate of $\boldsymbol{w}_l(\alpha)$ with the explanatory variables $\boldsymbol{x}_l$ and the resulting estimate is denoted by $\widehat{\boldsymbol{\Gamma}}_\alpha$. Finally the estimate $\widehat{\boldsymbol{B}}_\alpha$ of $\boldsymbol{B}$ is obtained by re-transforming $\widehat{\boldsymbol{\Gamma}}_\alpha$ by the matrix $\boldsymbol{E}_\alpha$ as

$$\widehat{\boldsymbol{B}}_\alpha = \widehat{\boldsymbol{\Gamma}}_\alpha \ \boldsymbol{E}_\alpha^{\ T}. \tag{3}$$

Since the estimate $\widehat{\boldsymbol{B}}_\alpha$ depends on the choice of $\boldsymbol{E}_\alpha$, it is essential to find the optimal subset index α^* based on some criterion. It has been dealt with in multivariate estimation problems by Chakraborty and Chaudhuri (1998), Chakraborty (1999). Depending on the nature of problems, there exist various criterions in describing the optimality. They used the criterion to minimize the generalized variance of the multivariate location or regression estimate. However, the asymptotic generalized variance of the estimate of $\widehat{\boldsymbol{B}}_\alpha$ depends on $\boldsymbol{E}_\alpha$ and it has a rather complex form. Thus it is nearly useless in general situations to calculate the generalized variance of some estimators.

One serious drawback of coordinatewise extension of univariate regression estimates in multivariate regression model is that such extensions do not take into account the inter-dependence that exists among the components of the response vector. It is a motive to suggest MLTS estimator. On the transformed data set $\{ \boldsymbol{x}_i, \ \boldsymbol{E}_\alpha^{-1} \ \boldsymbol{y}_i \}$ the multivariate regression model (1) can be rewritten as

$$\boldsymbol{w}_i(\alpha) = \boldsymbol{E}_\alpha^{-1} \ \boldsymbol{B}^T + \ \boldsymbol{e}_i^{\ *}, \tag{4}$$

where $\boldsymbol{e}_i^{\ *} = \boldsymbol{E}_\alpha^{-1} \ \boldsymbol{e}_i$. In model (4) we will get the coordinatewise LTS estimate $\widehat{\boldsymbol{\Gamma}}_\alpha$. To overcome the drawback of coordinatewise estimate $\widehat{\boldsymbol{\Gamma}}_\alpha$ the covariance matrix of $\boldsymbol{e}_i^{\ *}$ should be as orthogonal as possible in the d-dimensional vector space, that is $[\ \mathrm{cov}(\ \boldsymbol{e}_i^{\ *})]^{-1} = \boldsymbol{E}_\alpha^{\ T} \ \boldsymbol{\Sigma}^{-1} \ \boldsymbol{E}_\alpha = \lambda \boldsymbol{I}$. Hence we select α for which

$$\frac{\mathrm{trace}(\ \boldsymbol{E}_\alpha^{\ T} \ \boldsymbol{\Sigma}^{-1} \ \boldsymbol{E}_\alpha)}{|\ \boldsymbol{E}_\alpha^{\ T} \ \boldsymbol{\Sigma}^{-1} \ \boldsymbol{E}_\alpha|^{1/d}} \tag{5}$$

is minimized. Note that the above minimization problem (5) is equivalent to minimizing the ratio of the arithmetic and the geometric means of the eigenvalues of the matrix $\boldsymbol{E}_{\alpha}^{T}\ \boldsymbol{\Sigma}^{-1}\ \boldsymbol{E}_{\alpha}$. Also it is the same criterion as Chakraborty (1999).

Let $\boldsymbol{B}_I = \boldsymbol{X}_I^{-1}\ \boldsymbol{Y}_I$. Then $\boldsymbol{B}_I$ is the exact estimate of the regression model (1) based on the data set $\{(\boldsymbol{x}_{i_k}, \boldsymbol{y}_{i_k}), k=1,\ldots,p\}$. The estimate $\boldsymbol{B}_I$ will be appropriate if the residuals on whole data set are small. We adopt the concept of LTS regression for searching the optimal index set I from all possible subsets of size p of $\{1,\ldots,n\}$ as

$$\text{argmin}_I \sum_{k=1}^{h}[(\boldsymbol{y}_i - \boldsymbol{B}_I^T\ \boldsymbol{x}_i)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i - \boldsymbol{B}_I^T\ \boldsymbol{x}_i)]_{k:n} \tag{6}$$

where $a_{i:n}$ denotes the i-th order statistic from a set of a_i, $i=1,\ldots,n$. Here the value of h is called coverage. When $h=n$, the criterion (6) becomes the same as that of LS estimate.

Algorithm

(i) Obtain an affine equivariant and high breakdown estimate $\widehat{\boldsymbol{\Sigma}}$ of the scale matrix $\boldsymbol{\Sigma}$ of error vector $\boldsymbol{e}_i$ from $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$.

(ii) Choose I^* to satisfy (6). Given I^*, find J^* to minimize (5). Set $\alpha^* = I^* \cup J^*$.

(iii) Compute $\boldsymbol{E}_{\alpha^*}$ and transform response vector $\boldsymbol{y}_i$ to $\boldsymbol{w}_I = \boldsymbol{E}_{\alpha^*}\ \boldsymbol{y}_i$.

(iv) Obtain the coordinatewise LTS estimate $\widehat{\boldsymbol{\Gamma}}_{\alpha^*}$ on $\{(\boldsymbol{x}_i, \boldsymbol{w}_i)\}$. Then MLTS estimate $\widehat{\boldsymbol{B}}_{MLTS}$ becomes $\widehat{\boldsymbol{B}}_{\alpha^*} = \widehat{\boldsymbol{\Gamma}}_{\alpha^*}\ \boldsymbol{E}_{\alpha^*}^T$.

Note that while the transformed response vector $\widehat{\boldsymbol{\Sigma}}^{-1/2}\ \boldsymbol{y}$ in multivariate model (1) is a popular approach (Zellner, 1962), the transformation does not provide an affine equivalent modification of coordinatewise LTS estimate. The limitation of such an approach lies in the point that there does not exist an affine equivariant square root of usual estimates of the matrix $\boldsymbol{\Sigma}$ (Chakraborty, 1999).

We need an appropriate estimate of $\boldsymbol{\Sigma}$ to choose the optimal α^* satisfying (5) and (6). Here the estimate $\widehat{\boldsymbol{\Sigma}}$ should be affine equivariant and high breakdown for the statistical properties of MLTS estimator.

3. Properties of MLTS

In view of the definition of $\boldsymbol{B}_{\alpha}$ in (3), we have the following result, which asserts that the MLTS estimate is affine equivariant.

Proposition 1 *Let $\widehat{B}_\alpha(X, Y)$ be the estimate satisfying (3) on the data set (X, Y). The estimator $\widehat{B}_\alpha$ satisfies three version of equivariance in multivariate regression estimator (Rousseeuw et al., 2001).*

(a) *regression equivariance*

$$\widehat{B}_\alpha(X, XA+Y) = \widehat{B}_\alpha(X, Y)+A$$

(b) *y-affine equivariance*

$$\widehat{B}_\alpha(X, YU) = \widehat{B}_\alpha(X, Y)U$$

(c) *x-affine equivariance*

$$\widehat{B}_\alpha(XV, Y) = V^{-1}\,\widehat{B}_\alpha(X, Y)$$

for all nonsingular $d\times d$ matrices U, $p\times p$ matrices V and $p\times d$ matrices A.

Proof We will prove *(b)* only. Other properties can be proved in a similar manner. For y-affine equivariance, the i-th observation y_i is transformed to $U^T y_i$. Then Y_I and Y_J are transformed to $Y_I U$ and $Y_J U$, respectively. It means that E_α defined in (2) becomes $U^T E_\alpha$. The transformed response vector $w_l(\alpha)$ remains invariant under a non-singular linear transformation of y_l, because $(U^T E_\alpha)^{-1} U^T y_l = E_\alpha^{-1} y_l$. Thus the estimated matrix of regression parameters $\widehat{\Gamma}_\alpha$ obtained by regressing each coordinate of $w_l(\alpha)$ on x_l using LTS estimate separately is invariant under that transformation. Therefore $\widehat{B}_\alpha$ is transformed to $\widehat{\Gamma}_\alpha E_\alpha^T U$. It completes the proof ■

The coordinatewise LTS estimator is not y-affine equivariant and also it does not reflect the interrelationship among variables of error vectors. On the contrary the MLTS estimator is affine equivariant and it considers the correlations. The estimator MLTS satisfying (1) includes the scale matrix Σ, which is usually unknown. For affine equivariance of MLTS estimator the estimate $\widehat{\Sigma}$ should be affine equivariant.

Let us consider the global robustness of MLTS estimator. As a measure it is the finite-sample version of breakdown point, introduced by Donoho and Huber (1983). The breakdown point of an estimator $T(Z)$ at a sample Z is defined as

$$\varepsilon_n^*(T) = \min\{\frac{m}{n}; \sup\|T(Z) - T(\tilde{Z})\| = \infty\}$$

where $\tilde{Z}$ is obtained by replacing m observations by arbitrary points. Roughly speaking, the breakdown point is the smallest fraction of the contaminated data to make the estimate meaningless. When the sample size is n, the breakdown point of LS estimate is $1/n$. So the asymptotic breakdown point of LS estimate is 0.

It is apparent that the robustness of the MLTS estimate will critically depend on the robustness of the estimate $\widehat{\Sigma}$ used in its construction which could be seen in Section 2. The following proposition describes the breakdown property of MLTS estimate.

Proposition 2 *Let $Z=(X, Y)$ be a set of $n \geq p+d$ observations and $\widehat{\Sigma}$ of the scale parameter Σ with $\varepsilon_n^*(\widehat{\Sigma}) = \lceil n\gamma \rceil / n$ where $\gamma=(n-h)/n \leq (n-(p+d-1))/(2n)$. Assume also that observations are in general position. Then the finite sample breakdown point of MLTS estimate in regression model (1) satisfies $\varepsilon_n^*(\widehat{B}_{MLTS}) = \lceil n\gamma \rceil / n$. Consequently its asymptotic breakdown point is 50%.*

Proof Let $\widetilde{Z}$ be a data set obtained by replacing $m < \lceil n\gamma \rceil$ points from the original data set Z by arbitrary values. The estimate $\widehat{\Sigma}$ is a high breakdown estimate like the minimum covariance determinant estimate (Rousseeuw and Leroy, 1987). The index set α^* does not break, because $n > p+d$ and m is no more than $[n/2]$. Thus E_{α^*} will remain bounded. Furthermore each coordinatewise LTS estimate has breakdown point $([(n-p)/2]+1)/n$. Therefore $\min(\lceil n\gamma \rceil / n, ([(n-p)/2]+1)/n) = \lceil n\gamma \rceil / n$ completes the proof. ■

4. Simulation and Example

To investigate the performance of MLTS estimate in finite sample situations, we conducted a simulation study and analyzed a real data set for which there are some appropriate multi-response linear models.

4.1 Simulation

Simulation is conducted to compare the efficiency of the proposed MLTS estimate with coordinatewise LTS estimate. We consider the following multivariate regression model $y_i = B^T x_i + e_i$, where e_is are generated from bivariate normal distribution, bivariate Laplace distribution and bivariate t distribution with degrees of freedom 3 with the covariance matrix $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Here the length of y_i and x_i are all two, the first element of x_i is all one and the second element of x_i is generated from the standard univariate normal distribution. Using these e_i, x_i and $B = O$, we have generated the observations (x_i, y_i) for $i=1, \dots, n$. The efficiencies are computed using the fourth root of the ratio of the generalized variances of two competing estimates (Bickel, 1964) and the generalized variance is estimated using 1000 Monte Carlo replications. Here we obtained MLTS estimate with $h=[(3n)/4]$.

To illustrate the performance of MLTS estimate in the presence of correlation among response variables we simulate using the previous framework with sample size $n=20$ and 30. The efficiencies of MLTS estimate and coordinatewise LTS estimate are summarized in Table 1. It shows that the efficiency of MLTS estimate increases as correlation among coordinates of response vector increases. Thus we should consider the covariance matrix of error vector when we will obtain an estimate of regression coefficients in multivariate regression model. Table 1 shows that MLTS estimate appears to be more efficient than coordinatewise LTS estimate in regardless of error distributions.

4.2 Numerical example

A numerical example is given to illustrate the effectiveness of MLTS estimate. The data consist of 62 measurements of four pulp fibre characteristics (arithmetic fibre length, long fibre fraction, fines fibre fraction, zero span tensile) and the four paper properties (breaking length, elastic modulus, stress at failure, burst strength). The aim of analyzing the pulp fibre data is to predict four paper properties ($d=4$) from four fibre characteristics ($p=5$). See Lee (1992) and Rousseeuw et al. (2001) for detail.

Table 1. Estimated efficiencies of MLTS estimates with respect to coordinatewise LTS estimates when error distribution comes from bivariate normal, bivariate Laplace and bivariate t with degrees of freedom 3.

Error distribution	Sample size	ρ				
		0.2	0.5	0.7	0.8	0.95
Normal	20	103%	114%	124%	149%	198%
	30	105%	114%	137%	147%	198%
Laplace	20	103%	108%	135%	148%	212%
	30	110%	114%	134%	146%	210%
t	20	106%	111%	141%	152%	198%
	30	107%	116%	133%	149%	211%

First we applied classical multivariate regression to this data. Figure 1 represents the results of classical analysis, which depicts the Mahalanobis distances of residuals versus the Mahalanobis distances of predictors. The value of y-axis is $\sqrt{\boldsymbol{r}_{i,LS}^T \hat{\boldsymbol{\Sigma}}_{LS}^{-1} \boldsymbol{r}_{i,LS}}$ where $\boldsymbol{r}_{i,LS} = \boldsymbol{y}_i - \hat{\boldsymbol{B}}_{LS}^T \boldsymbol{x}_i$ and $\hat{\boldsymbol{\Sigma}}_{LS}$ denotes the sample covariance matrix of LS residuals. The value of x-axis is $\sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \hat{\boldsymbol{\Sigma}}_{xx}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})}$ where $\bar{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{xx}$ are the sample mean and sample covariance of $\boldsymbol{x}_i$, respectively. This plot gives the information about regression outliers (y-axis) and leverage points (x-axis). The horizontal and vertical cut off lines are all $\sqrt{\chi^2_{4,0.975}} = 3.34$. From Figure 1 we observe that observations 60, 61 are leverage points and observations 51, 52, 56 are regression outliers.

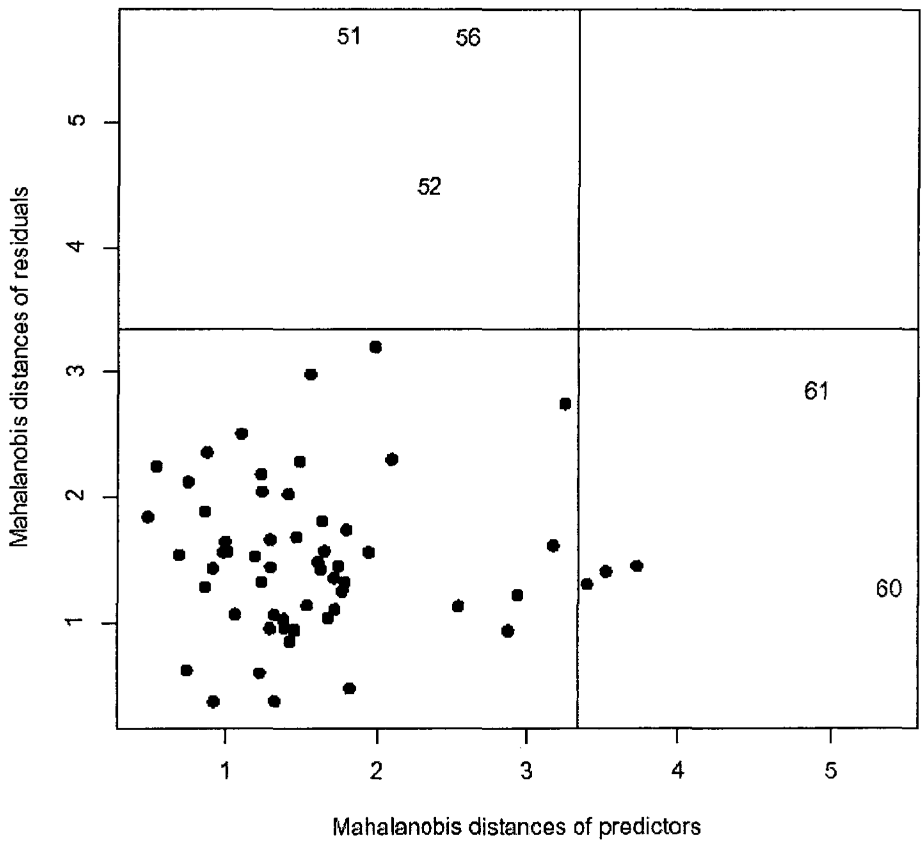

Figure 1. Plot of Mahalanobis distances of LS residuals versus Mahalanobis distances of predictors.

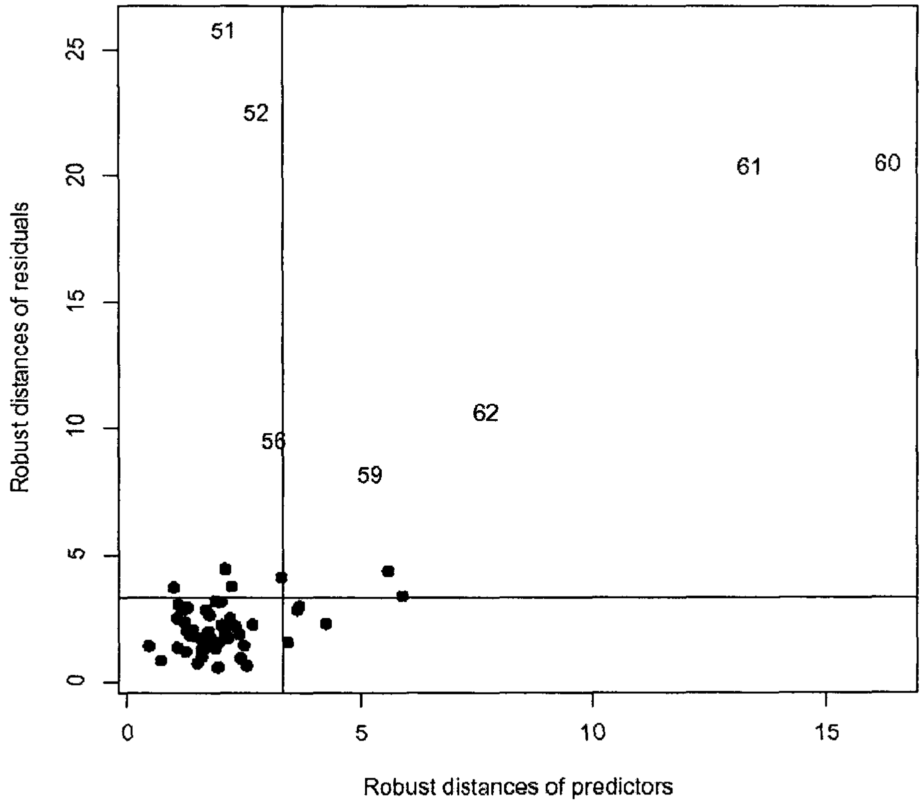

Figure 2. A diagnostic plot of robust distances of residuals versus robust distances of predictors.

Next we performed MLTS estimate with $h=46$. Figure 2 depicts a diagnostic plot of robust distances of residuals versus robust distances of predictors (Rousseeuw and Zomeren, 1990). In this figure the value of y-axis is $\sqrt{r_{i,MLTS}^{T}\,\widehat{\Sigma}^{-1}\,r_{i,MLTS}}$, where $r_{i,MLTS}=y_i-\widehat{B}_{MLTS}^{T}x_i$ and that of x-axis is $\sqrt{(x_i-\widehat{\mu}_x)^T\,\widehat{\Sigma_{xx}}^{-1}(x_i-\widehat{\mu}_x)}$, where $\widehat{\mu}_x$ and $\widehat{\Sigma}_{xx}$ are the minimum covariance determinant estimates of mean vector and covariance matrix for the data matrix X, respectively. And the horizontal and vertical cutoff lines are the same as that of Figure 1. Figure 2 shows that observations 51, 52, 56, 59, 60, 61, 62 are regression outliers and observations 59, 60, 61, 62 are leverage points. We may conclude that observations 51, 52, 56, 59, 60, 61, 62 are bad points, because they are separated from others. By the report on collecting data, all except observations 59-61 were produced from fir wood. Moreover, outlying observations were obtained using different pulping process. For example observations 62 is the only sample from a chemi-thermomechanical pulping process and observations 60, 61 are the only samples from a solvent pulping process. Finally observations 51, 52, 56 are obtained from a kraft pulping process (Rousseeuw, et al., 2001 ; Lee, 1992). Figure 2 explains well the history of the pulp fibre data set. However, classical multivariate regression only detected observations 51, 52, 56 of these outliers and considered observations 60, 61 to be good leverage points. It shows that MLTS estimate provides the useful information on detecting outliers in a multivariate regression problem.

References

[1] Bickel, P. J. (1964). On some alternative estimates for shift in the p-variate one sample problem, *Annals of Mathematical Statistics*, 35, 1079-1090.

[2] Chakraborty, B. (1999). On multivariate median regression, *Bernoulli*, 5, 683-703.

[3] Chakraborty, B. and Chaudhuri, P. (1998). On an adaptive transformation and retransformation estimate of multivariate location, *Journal of the Royal Statistical Society Series B*, 60, 145-157.

[4] Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point, In *A Festschrift for Erich Lehmann*, Edited by P. Bickel, K. Doksum and J. L. Hodges, 157-184, Wadsworth, Belmont, CA.

[5] Hampel, F. R., Ronchetti, E. M. and Rousseeuw, P. J. (1986). *Robust Statistics : The Approach Based on Influence Functions*, John Wiley, New York.

[6] Lee, J. (1992). *Relationships between Properties of Pulp-Fibre and Paper*, Ph. D. thesis, University of Toronto.

[7] Ollila, E., Oja, H. and Hettmansperger, T. P. (2002). Estimates of regression coefficients based on the sign covariance matrix, *Journal of the Royal Statistical Society Series B*, 64, 447-466.

[8] Rao, C. R. (1988). Methodology based on the L_1-norm in statistical inference, Sankhyā, 50, 289-313.

[9] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*, John Wiley, New York.

[10] Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, 85, 633-651.

[11] Rousseeuw, P. J., Van Alest, S. and Van Driessen, K. (2001). Robust multivariate regression, *To be published.*

[12] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association*, 57, 348-368.

[Received August 2003, Accepted October 2003]