The Korean Communications
in Statistics Vol. 12 No. 1, 2005
pp. 39-46

A Robust Estimator in Multivariate Regression Using Least Quartile Difference

Kang-Mo Jung[1]

Abstract

We propose an equivariant and robust estimator in multivariate regression model based on the least quartile difference (LQD) estimator in univariate regression. We call this estimator as the multivariate least quartile difference (MLQD) estimator. The MLQD estimator considers correlations among response variables and it can be shown that the proposed estimator has the appropriate equivariance properties defined in multivariate regressions. The MLQD estimator has high breakdown point as does the univariate LQD estimator. We develop an algorithm for MLQD estimate. Simulations are performed to compare the efficiencies of MLQD estimate with coordinatewise LQD estimate and the multivariate least trimmed squares estimate.

Keywords : Breakdown point; Equivariance; Least quartile difference estimator; Multivariate regression; Outliers.

1. Introduction

Consider the linear regression model given by

$$y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i, \qquad 1 \le i \le n$$

where $\boldsymbol{\beta}$ is the p-dimensional parameter including the intercept. The residuals are denoted by $r_i(\widehat{\boldsymbol{\beta}}) = y_i - \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. The least squares (LS) estimator which minimizes the sum of the squared residuals is most well known, because it is simple and has the closed form solution to a certain system 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, but the breakdown points of these estimators cannot exceed the inverse of the dimension of

[1] Associate professor, Department of Informatics & Statistics, Kunsan National University, Kunsan 573-701, Korea.
Email: kmjung@kunsan.ac.kr

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. Croux et al. (1994) proposed the least quartile difference (LQD) estimator, which minimizes the lower quartile of the ordered absolute differences in residual pairs, that is,

$$\hat{\beta}_{LQD} = \text{argmin}_{\beta} \{ |r_i - r_j|;\ i < j \}_{\binom{h_p}{2}:\binom{n}{2}},$$

where $h_p = [(n+p+1)/2]$, p is the number of regression parameters, and the notation $\binom{h_p}{2}:\binom{n}{2}$ means minimize the $\binom{h_p}{2}$th order statistic among $\binom{n}{2}$ elements of the set $\{ |r_i - r_j|;\ i < j \}$. The LQD estimator has a 50% breakdown point and is asymptotically normal with Gaussian efficiency of 67%, whereas LMS has asymptotic efficiency of 0% and LTS has asymptotic efficiency of only 8%. 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 each response. Chakraborty (1999) suggested a new extension of least absolute deviation regression based on the so-called transformation and retransformation method. Also 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. Jung (2003) proposed the multivariate LTS (MLTS) estimator based on LTS. As pointed in the previous paragraph, the LTS estimator has a low asymptotic efficiency when the errors are generated from a univariate normal distribution, and so does the MLTS estimator.

In this paper we propose an affine and robust estimator of regression parameters in multivariate linear regression model. This estimator is based on the LQD estimator in univariate regression model, which is a most high breakdown estimator and high efficient rather than LMS and LTS. 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, 1999) for regression equivariance and it also uses a covariance matrix of error vectors. The proposed estimator has 50% breakdown point, because it inherits the breakdown point of LQD.

In Section 2 we define a multivariate regression model and propose a new estimator. We call this estimator as the multivariate least quartile difference (MLQD) estimator. We develop an algorithm to compute MLQD 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 is given to illustrate the efficiency of our proposed estimate. Simulation results show that the MLQD estimate appears to be more efficient than coordinatewise LQD estimate when there exist correlations among variables of error vector.

2. Multivariate Least Quartile Difference Estimator

Consider the multivariate linear regression model

$$y_i = B^T x_i + e_i \; ; \qquad i = 1, \cdots, n, \tag{1}$$

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

Consider n data points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, and assume that $n > d+p$. Write $\alpha = \{i_1, i_2, \cdots, i_p, j_1, \cdots, j_d\}$ and let $W(\alpha)$ be the $p \times p$ matrix whose k-th column vector is x_{i_k}, and $Z(\alpha)$ be the $d \times p$ matrix whose k-th column vector is y_{i_k}. We will assume that $W(\alpha)$ is invertible and define $E(\alpha)$ to be the $d \times d$ matrix consisting of the columns

$$Y_{j_1} - Z(\alpha)\, W(\alpha)^{-1} x_{j_1}, \cdots, y_{j_d} - Z(\alpha)\, W(\alpha)^{-1} x_{j_d}. \tag{2}$$

If the error vectors e_i's are i.i.d. random vectors with a common probability distribution, which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, the matrix $E(\alpha)$ will be invertible with probability one. We define the transformation response vectors as $w_l(\alpha) = E(\alpha)^{-1} y_l$ for $1 \le l \le n$ and $l \notin \alpha$. We apply univariate LQD regression on each coordinate of $w_l(\alpha)$ with the explanatory variables x_l and the resulting estimate is denoted by $\widehat{\Gamma}(\alpha)$. Finally the estimate $\widehat{B}(\alpha)$ of B is obtained by re-transforming $\widehat{\Gamma}(\alpha)$ by the matrix $E(\alpha)$ as

$$\widehat{B}(\alpha) = \widehat{\Gamma}(\alpha)\, E(\alpha)^T. \tag{3}$$

Since the estimate $\widehat{B}(\alpha)$ depends on the choice of $E(\alpha)$, 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 (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 $\widehat{B}(\alpha)$ depends on $E(\alpha)$ 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 the MLQD estimator. On the transformed data set $\{x_i, E(\alpha)^{-1} y_i\}$ the multivariate regression model (1) can be rewritten as

$$w_i(\alpha) = E(\alpha)^{-1} B^T + e_i^*, \tag{4}$$

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

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

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 $E(\alpha)^T \Sigma^{-1} E(\alpha)$. Also it is the same criterion as Chakraborty (1999).

Let $B_I = X_I^{-1} Y_I$. Then B_I is the exact estimate of the regression model (1) based on the data set $\{(x_{i_k}, y_{i_k}), k=1,\cdots,p\}$. The estimate B_I will be appropriate if the residuals on whole data set are small. We adopt the concept of LMS regression for searching the optimal index set I from all possible subsets of size p of $\{1,\cdots,n\}$ as

$$\operatorname{argmin}_I[(y_i - B_I^T x_i)^T \widehat{\Sigma}^{-1}(y_i - B_I^T x_i)]_{h:n} \tag{6}$$

where $a_{i:n}$ denotes the i-th order statistic from a set of a_i, $i=1,\cdots,n$. Here the value of h is called coverage.

Algorithm

(i) Obtain an affine equivariant and high breakdown estimate $\widehat{\Sigma}$ of the scale matrix Σ of error vector e_i from $\{(x_i, y_i)\}$.

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

(iii) Compute $E(\alpha^*)$ and transform response vector y_i to $w_i = E(\alpha^*)\, y_i$.

(iv) Obtain the coordinatewise LQD estimate $\widehat{\Gamma}(\alpha^*)$ on $\{(x_i, w_i)\}$. Then MLQD estimate

$\widehat{\boldsymbol{B}}_{MLQD}$ becomes $\widehat{\boldsymbol{\Gamma}}(\alpha^*)\ \boldsymbol{E}(\alpha^*)^T$.

Note that while the transformed response vector $\widehat{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{y}$ in multivariate model (1) is a popular approach, the transformation does not provide an affine equivalent modification of coordinatewise LQD 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). Here the estimate $\widehat{\boldsymbol{\Sigma}}$ should be affine equivariant and high breakdown for the statistical properties of MLQD estimator.

3. Properties of MLQD

In view of the definition of $\boldsymbol{B}_\alpha$ in (3), we have the following result, which asserts that the MLQD estimate is affine equivariant. See Jung (2003) for proof.

Proposition 1 *Let* $\widehat{\boldsymbol{B}}(\alpha)$ *be the estimate satisfying (3) on the data set* $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_n, \boldsymbol{y}_n)$. *The estimator* $\widehat{\boldsymbol{B}}(\alpha)$ *satisfies*

(a) *affine equivariance : Suppose that* $\boldsymbol{A}$ *is a* $d\times d$ *nonsigular matrix. Then the estimator based on* $(\boldsymbol{x}_1, \boldsymbol{A}\ \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{A}\ \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_n, \boldsymbol{A}\ \boldsymbol{y}_n)$ *is given by* $\boldsymbol{A}\ \widehat{\boldsymbol{B}}(\alpha)$

(b) *regression equivariance : Suppose that the response vectors,* $\boldsymbol{y}_i$*'s, are transformed to* $\boldsymbol{y}_i - \boldsymbol{G}\ \boldsymbol{x}_i$ *for* $i = 1, \cdots, n$, *where* $\boldsymbol{G}$ *is a fixed* $d\times p$ *matrix. Then the estimator will be transformed to* $\widehat{\boldsymbol{B}}(\alpha) - \boldsymbol{G}$.

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

Let us consider the global robustness of MLQD 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 $\boldsymbol{T}(Z)$ 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 MLQD 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 MLQD 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 MLQD estimate in regression model (1) satisfies* $\varepsilon_n^*(\widehat{B}_{MLQD}) = \lceil n\gamma \rceil /n$. *Consequently its asymptotic breakdown point is 50%.*

Proof Let $\tilde{Z}$ 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 $E(\alpha^*)$ will remain bounded. Furthermore each LQD estimate for the component of the response vector has breakdown point $([(n-p)/2]+1)/n$ (Croux et al. (1994)) where $h_p = [(n+p+1)/2]$. Therefore $\min(\lceil n\gamma \rceil /n, ([(n-p)/2]+1)/n) = \lceil n\gamma \rceil /n$ completes the proof. ■

4. Simulation

To investigate the performance of the MLQD estimate in finite sample situations, we conducted a simulation study. Simulation is conducted to compare the efficiency of the proposed MLQD estimate with coordinatewise LQD estimate. We consider the following multivariate regression model $y_i = B^T x_i + e_i$, where e_i's 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 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for $i = 1, \cdots, n$. The proposed MLQD estimates are compared with the LS estimates, the MLTS estimate (Jung, 2003) and the coordinatewise LQD estimate, and for the purpose of efficiency computation, the estimates of their generalized variances were obtained based on 1000 Monte Carlo replications. The relative efficiencies are taken to be the fourth root of the ratio of the generalized variances of two competing estimates of $\boldsymbol{B}$ (Bickel, 1964).

To illustrate the performance of the MLQD estimate in the presence of correlation among response variables we simulate using the previous framework with sample size $n = 20, 30, 40$ and 50. For $n = 40$ and 50 we obtained results similar to the case $n = 20$ and 30. For the latter the efficiencies of the MLQD estimates and the coordinatewise LQD estimates are summarized in Table 1. It shows that the efficiency of the MLQD 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. On this point the MLQD estimator considers the covariance matrix of error vectors. Table 1 shows that the MLQD estimate appears to be more efficient than the coordinatewise LQD estimate in the presence of substantial correlations in regardless of error distributions.

In order to compare the efficiencies of the MLQD estimates and the MLTS estimates, we performed the simulations using the same framework with the previous work. Table 2 presents the efficiencies of the MLQD estimates and the MLTS estimates. We observe that the MLQD estimates are more efficient than the MLTS estimates when the errors are generated from the multivariate normal. When the tail probabilities are larger, the efficiencies of two estimates are similar. However, we see that the MLQD estimate is more efficient than the MLTS estimate.

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

Error distribution	Sample size	ρ		
		0.0	0.5	0.9
Normal	20	0.98	1.04	1.40
	30	0.98	1.07	1.37
Laplace	20	0.96	0.97	1.33
	30	0.96	1.04	1.41
t	20	0.93	0.98	1.45
	30	0.96	1.04	1.48

Table 2. Estimated efficiencies of the MLQD estimates with respect to the MLTS estimates when error distribution comes from bivariate normal, bivariate Laplace and bivariate t with degrees of freedom 3.

Error distribution	Sample size	ρ		
		0.0	0.5	0.9
Normal	20	1.27	1.22	1.27
	30	1.45	1.46	1.43
Laplace	20	1.15	1.03	1.18
	30	1.04	1.12	1.07
t	20	1.15	1.11	1.13
	30	0.99	0.98	1.03

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] Croux, C., Rousseeuw, P. J. and Hössjer, O. (1994). Generalized S-estimators, *Journal of the American Statistical Association*, 89, 1271-1281.

[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] Jung, K.-M. (2003). An equivariant and robust estimator in multivariate regression based on least trimmed squares, *The Korean Communications in Statistics*, 10, 1037-1046.

[6] 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.

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

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

[Received May 2004, Accepted November 2004]