On Confidence Intervals of High Breakdown Regression Estimators

Dong-Hee Lee[*], YouSung Park[†] and Kang-yong Kim[‡]

Abstract

A weighted self-tuning robust regression estimator (WSTE) has the high breakdown point for estimating regression parameters such as other well known high breakdown estimators. In this paper, we propose to obtain standard quantities like confidence intervals, and it is found to be superior to the other high breakdown regression estimators when a sample is contaminated.

Keywords : High breakdown point, Confidence interval, Robust regression, Outlier

1 Introduction

A regression estimator is said to be robust if it is still reliable in the presence of outliers. Moreover, a robust regression method can be evaluated by two concepts of robustness: local robustness and global robustness (Hernandez and Yohai 2003). The local robustness of an estimator refers to its stability as the fraction of outliers in the data tends to zero, whereas global robustness concerns the stability of the estimator when a large fraction of outliers is included in the sample. Local robustness can be measured by the influence function (Hampel 1974) which indicates how much an estimator is influenced by a single outlying observation, while global robustness can be measured by the breakdown point (Hampel 1971, Donoho and Huber 1983) which is the maximum fraction of outliers that renders an estimator useless.

High breakdown regression estimators are developed to give the highest breakdown point 50%. Indeed, 50% is the best that can be expected because it becomes impossible to distinguish between the *good* and the *bad* parts of the sample. The estimators, Least Median of Squares (LMS) by Hampel (1975), Least Trimmed Squares (LTS) by Rousseeuw (1984), and S-estimators by Rousseeuw

[*] Dong-Hee Lee is Post-doc, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-gu, KOREA

[†] YouSung Park is Professor, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-gu, KOREA

[‡] Kang-yong Kim is Graduate student, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-gu, KOREA.

and Yohai (1984), have the highest breakdown point of 50%. Other high breakdown regression estimators included in this setup are τ-estimators (Yohai and Zamar 1988), R-estimators (Hössjer 1994), GS-estimators such as the least quartile difference (LQD) and the least trimmed difference regression estimator (LTD) (Croux, Rousseeuw and Hössjer 1994, Stromberg, Hössjer and Hawkins 2000) which are based on differences of residuals and indicated to be more efficient than other high breakdown estimators.

By implication, although these high breakdown regression estimators have different criteria with each others, they accommodate the best fitting cases chosen by their own criteria. Moreover, computing the high breakdown estimators is notoriously difficult in large sample and even in small sample for practically obtaining high breakdown regression estimates. To avoid this problem, high breakdown estimators typically use only a subset of the data which is randomly selected by a resampling technique. This resampling technique renders the high breakdown estimators that do not have the permutation invariance. This invariance is another desired property for an estimator to be invariant under any permutation of observations. The resampling causes a lower convergence rate and a lower breakdown point than theoretically expected (Hawkins and Olive 2002). In contrast, the weighted self-tuning robust regression estimator (WSTE) use resampling technique no more even though it has the high breakdown point 50% such as other high breakdown estimators (Lee 2004).

In this paper, we propose the method in order to obtain standard quantities like t-value, confidence intervals, and like this, and investigate the performance of WSTE in constructing confidence intervals of regression coefficients while the presence of outliers. The proposed method is presented in Section 2. Result of a simulation study carried out for evaluating its performance over other high breakdown regression estimators are presented in Section 3.

2 The Weighted Self-Tuning Robust Regression Estimator

We consider the linear regression model given by

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i, \quad 1 \leq i \leq n,$$

where $\boldsymbol{\beta}$ is the p-dimensional parameter including an intercept parameter. The random sample $\{\mathbf{x}_i, y_i\}$ are from a continuous distribution F and the error ϵ has a distribution F_ϵ with mean 0 and a finite variance σ^2, which is independent with $\mathbf{x}_i$. Let $\tilde{\boldsymbol{\beta}}_n$ be a trial high breakdown estimator. If regressor coefficients $\tilde{\boldsymbol{\beta}}_n$ lead to a linear model fitting well with the data, residuals using $\tilde{\boldsymbol{\beta}}_n$ are small for the majority of the data while large for outliers. To find such a $\tilde{\boldsymbol{\beta}}_n$, we first define a collection of subsets of n observations as follows. Based on response variable y, let $\tilde{y}_{q_1}$, $\tilde{y}_{q_2}$, and $\tilde{y}_{q_3}$

be the first quartile, the median, and the third quartile of y, respectively. Then define

$$O_{01} = \{(\mathbf{x}_j, y_j) | \tilde{y}_{q_1} \leq y_j < \tilde{y}_{q2}\} \text{ and } O_{02} = \{(\mathbf{x}_j, y_j) | \tilde{y}_{q_2} \leq y_j \leq \tilde{y}_{q3}\}.$$

Next, based only on the ith independent variable, partition n observations into four quadrants. To that end, set

$$\bar{y}_{u_i} = \left(\sum_{j=1}^{n} y_j I[x_{ij} \geq \bar{x}_i] \right) / \left(\sum_{j=1}^{n} I[x_{ij} \geq \bar{x}_i] \right)$$

and

$$\bar{y}_{l_i} = \left(\sum_{j=1}^{n} y_j I[x_{ij} < \bar{x}_i] \right) / \left(\sum_{j=1}^{n} I[x_{ij} < \bar{x}_i] \right)$$

where $\bar{x}_i = 1/n \sum_{j=1}^{n} x_{ij}$. Then, for $i = 1, \cdots, p$ where p is the number of independent variables, define

$$O_{i1} = \{(\mathbf{x}_j, y_j) \mid x_{ij} \geq \bar{x}_i \text{ and } y_j \geq \bar{y}_{u_i}\},\ O_{i2} = \{(\mathbf{x}_j, y_j) \mid x_{ij} \geq \bar{x}_i \text{ and } y_j < \bar{y}_{u_i}\},$$
$$O_{i3} = \{(\mathbf{x}_j, y_j) \mid x_{ij} < \bar{x}_i \text{ and } y_j \geq \bar{y}_{l_i}\}, \text{ and } O_{i4} = \{(\mathbf{x}_j, y_j) \mid x_{ij} < \bar{x}_i \text{ and } y_j < \bar{y}_{l_i}\},$$

where we set $O_{ij} = \phi$ for all $i = 1, \cdots, p$ and $j = 1, \ldots, 4$ when any x_{ij} or y_j is infinite. Let $\mathbf{O}_i$ be the closure of the class $\{O_{ij} : 1 \leq j \leq 4\}$ under the union operation, i.e.

$$\mathbf{O}_i = \{O_{i1}, \cdots, O_{i4}; O_{i1} \cup O_{i2}, \cdots, O_{i3} \cup O_{i4}; O_{i1} \cup O_{i2} \cup O_{i3}, \cdots, O_{i2} \cup O_{i3} \cup O_{i4}; \bigcup_{k=1}^{4} O_{ik}\},$$

and $\mathcal{C} = \bigcup_{i=0}^{p} \mathbf{O}_i$ where $\mathbf{O}_0 = \{O_{01}, O_{02}, O_{01} \cup O_{02}\}$. Note that there are at most $14p + 4$ non-empty different sets in $\mathcal{C}$. Let K denote the number of sets in $\mathcal{C}$ having cardinality at least $p + 2$ and further let $E_1, E_2, \cdots, E_K$, $K \leq 14p + 4$ be an enumeration of these "thick" sets. We refer to an E_i as an elementary set for $i = 1, \cdots, 14p + 4$.

Using the elementary sets, the following four steps precisely describe how to obtain an optimal $\tilde{\beta}_n$ and the WSTE.

Step 1 Obtain OLS estimate $\mathbf{b}_k$ from observations in E_k. With $\mathbf{b}_k$ as regressor coefficients calculate for all n data points standardized residuals $\tilde{r}_j(\mathbf{b}_k) = r_j(\mathbf{b}_k)/s(\mathbf{b}_k)$ where $r_j(\mathbf{b}_k) = y_j - \mathbf{x}_j^t \mathbf{b}_k$, $j = 1, 2, \cdots, n$ and $s(\mathbf{b}_k) = (0.6745)^{-1}\text{median}(|r_j(\mathbf{b}_k)|)$. Do this for each of the reduced samples $E_1, E_2, \cdots, E_K$ and then obtain

$$\tilde{\beta}_n = \arg\min_{\mathbf{b}_k} \sum_{|\tilde{r}_j(\mathbf{b}_k)| < c_1, |\tilde{r}_i(\mathbf{b}_k)| < c_1} \left[r_i^2(\mathbf{b}_k) - r_j^2(\mathbf{b}_k)\right]_+$$

where c_1 is a cut-off value, $k = 1, 2, \cdots, K$, $[x]_+ = \max(0, x)$. Here the summation is taken over all i, $j = 1, 2, \cdots, n$. Define $O(\tilde{\beta}_n)$ to be the observations satisfying $|\tilde{r}_j(\tilde{\beta}_n)| < c_1$ for $j = 1, 2, \cdots, n$.

Step 2 Only using observations in $O(\tilde{\beta}_n)$, calculate a preliminary self-tuning estimator (PSTE) $T_{PSTE}(F_n)$ satisfying

$$\sum_{i=1}^{n} r_i \lambda(r_i) \mathbf{x}_i = \mathbf{0}.$$

where

$$\lambda(r_i) = \left(1 + \frac{n^{1-R_n^2} \sum_{j=1}^{n} \left[r_i^2 - r_j^2 \right]_+}{\sum_{j=1}^{n} \sum_{k=1}^{n} \left[r_j^2 - r_k^2 \right]_+} \right)^{-1}.$$

Step 3 Remove the observations with $|\tilde{r}_j(T_{PSTE}(F_n))| > c_2$ for $j = 1, 2, \cdots, n$ where $c_2 \leq c_1$. We call the removed observations temporary outliers. When no temporary outlier is detected or the remaining number of observations is less than $p + 2$, the $T_{PSTE}(F_n)$ in Step 2 is our WSTE. Otherwise, denote the remaining observations by $\mathbb{S}_{n_1}$ where n_1 is the size of $\mathbb{S}_{n_1}$.

Step 4 Based only on $\mathbb{S}_{n_1}$, construct new E_k's and then repeat **Step 1** and **Step 2**.

The estimate from Step 4 is the WSTE denoted by $T_{WSTE}(F_n)$. Observations with $|\tilde{r}_j(T_{WSTE}(F_n))| > c_2$ for $j = 1, 2, \cdots, n$ are deemed to be outliers.

The basic principle of WSTE is to fit the majority of the data, after which outliers may be identified as those points that lie far away from the fit, that is, the cases with large positive or large negative residuals. Therefore, we can apply a weighted least squares analysis based on the identification of the outliers following Rousseeuw and Leroy (1987) or construct a hybrid type estimator with the initial estimates using this WSTE such as Yohai (1987), Simpson, Ruppert and Carroll (1992), and Coakly and Hettmansperger (1993). Although it is well known that these one-step estimators can improve the rate of the initial estimator, the performance of a one step estimator when applied to an incosistenct high breakdown regression estimator to be an open question (Olive 2003). Therefore, we consider the weighted least squares based on the identification by the following weights:

$$w_i = \begin{cases} 1 & \text{if } |\frac{r_i}{\hat{\sigma}}| \leq 3 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This means simply that case i will be retained in the weighted least squares if its WSTE residual is small to moderate, but disregarded if it is an outlier. (1) is the *smooth* resection rule by Rousseeuw and Leroy (1987).

3 Comparisons

A simulation study is conducted to compute the empirical coverage probabilities of confidence intervals for the proposed estimate in Section 2. The multiple regression model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_5 x_{5i} + \epsilon_i,$$

where $i = 1, \cdots, n$, is considered. Simulations generate a fixed percentage of clean observations and clean observations are generated as follows. Each values of the independent variables for the clean observations are generated identically and independently from a normal distribution with a mean $\mu_x = 7.5$ and standard deviation of $\sigma_x = 4.0$. The response for the ith clean observation is generated by (2) where all βs are set to be 1 and ϵ_i is the random error distributed by $N(0,1)$. We consider the sample size $n = 50$ and outlier percentages from 0% to 40% by increasing 10%. After the clean observations are generated, the remainder of whole sample is filled with outliers. $x_{ij} = \bar{x}_{i,clean} + 5\sigma_x + \mathrm{N}(0, \sigma_x^2)$ and $y_j = \bar{y}_{clean} + \mathrm{N}(0,1)$ where $\bar{y}_{clean}$ is the sample mean of clean y's. Thus, it describes bad leverage outlying observations located near the centroid of y,

Table (1) shows the coverage probabilities of 7 regression estimators at 1000 simulation runs. Each cell indicates the mean of coverage probabilities of 5 regression coefficients. S and WSTE estimators are nearly exact comparable to OLS when a sample has no outlier. Moreover, S estimator shows the best performance while the sample is moderate in contamination level, but it is broken down over 20%. LMS and LTS have the similar result comparable to their generalized estimators, LQD and LTS. But LMS and LTS are better than LQD and LTS at 40% contamination level. On the other hand, the WSTE does not manifest any of the drawbacks which were noted in other estimators.

Table 1: Means of Coverage Probabilities of 95% and 99% Confidence Intervals

Nominal	Contamination	Estimators						
Level		OLS	LMS	LTS	S	LQD	LTD	WSTE
	0.0	95.1	73.7	63.9	93.5	73.9	62.6	92.4
	0.1		81.1	74.2	94.1	85.3	76.4	93.2
95%	0.2		88.3	83.0	94.4	92.0	88.4	93.8
	0.3		91.6	88.9	3.6	92.6	94.0	94.3
	0.4		71.9	75.4	0.9	2.2	44.0	94.8
	0.0	98.7	85.5	77.2	97.9	85.9	73.9	97.8
	0.1		92.0	86.2	98.6	93.7	87.4	98.2
99%	0.2		95.5	92.5	98.6	97.6	95.4	98.4
	0.3		97.7	96.4	9.8	97.0	98.9	98.9
	0.4		77.5	81.4	2.9	6.8	3.6	98.7

4 Discussion

The high breakdown regression estimators suffer from the computational problem. To overcome the computational problem, a resampling technique is generally adapted. Because of resampling, they are calculated using only partial observations and this may damage the robustness theoretically expected. On the other hand, No computational problem arises in calculating the WSTE. This is achieved by partitioning the observations into a finite number of subsets based on the means of independent and dependent variables. Therefore, WSTE is found to be superior to other high breakdown regression estimators in constructing confidence intervals of regression coefficients in the presence of outliers.

References

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

[2] Hawkins, D.M. and Olive, D.J. (2002). Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm. *Journal of the American Statistical Association*, **97**, 136-148.

[3] Hernandez S. and Yohai, V.J. (2003). Combining locally and globally robust estimates for regression. *Journal of Statistical Planning and Inference.* **113**, 633-661.

[4] Lee, Dong-Hee (2004). *Slef-Tuning Robust Regression Estimator.* Unpublished Ph.D. Thesis, Department of Statistics, Korea University.

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

[6] Rousseeuw, P.J. and Van Driessen, K. (1998). Computing LTS regression for large data sets. Technical Report, University of Antwerp, submitted.

[7] Simpson, D.G., Ruppert, D. and Carroll, R.J. (1992). On one-step GM estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*, **87**, 439-450.

[8] Stromberg, A.J., Hössjer, O. and Hawkins, D.M. (2000). The trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association*, **95**, 853-864.

[9] Yohai, V.J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, **15**, 642-656.