Self-tuning Robust Regression Estimation

YouSung Park*, DongHee Lee†

Abstract

We introduce a new robust regression estimator, self-tuning regression estimator. Various robust estimators have been developed with discovery for theories and applications since Huber introduced M-estimator at 1960's. We start by announcing various robust estimators and their properties, including their advantages and disadvantages, and furthermore, new estimator overcomes drawbacks of other robust regression estimators, such as ineffective computation, on preserving robustness properties.

Keywords : High breakdown point, Bounded influence, Robust regression estimation

1 Introduction

One of the most important statistical tools is a linear regression analysis for many fields. Nearly all regression analysis relies on the method of least squares for estimation of the parameters in the model. A problem that we often encountered in the application of regression is the presence of an outlier or outliers in the data. Outliers can be generated by from a simple operational mistake to including small sample from a different population, and they make serious effects of statistical inference. Even one outlying observation can destroy least squares estimation, resulting in parameter estimates that do not provide useful information for the majority of the data.

Robust regression analysis have been developed as an improvement to least squares estimation in the presence of outliers and to provide us information about what a valid observation is and whether this should be thrown. The primary purpose of robust regression analysis is to fit a model which represent the information in the majority of the data.

The properties of efficiency, breakdown point, and bounded influence are used to define the measure of robust technique performance in a theoretical sense. Efficiency can tell us how well a

*YouSung Park is a professor, Department of Statistics, Korea University, 5-1 Anam-Dong, Sungbuk-gu, KOREA

†DongHee Lee is a graduate student, Department of Statistics, Korea University, 5-1 Anam-Dong,Sungbuk-gu, KOREA.

robust technique performs relative to least squares on clean data (without outliers). High efficiency is mostly desired on estimation. The breakdown point is a measure for stability of the estimator when the sample contains a large fraction of outliers(Hampel,1974; Donoho and Huber,1983). It gives the minimum fraction of outliers which may produce an infinite bias. It is referred as the measure of global robustness in this sense.

For example, least squares has a breakdwn point of $1/n$. This indicates that only one outlier can make the estimates useless. In contrast, some robust regression estimates attaches approximately 50% breakdown point, and it is called a high breakdown point in this case.

Lastly, bounded influence is designed to counter the tendency of least squares to allow exterior **X**-space or high leverage points to exhibit greater influence, which can be especially important if these points are outliers.

Robust regression estimators were first introduced by Huber(1973,1981), and it is well known as M-regression estimator. But this estimator is not robust in the view point that it has the same breakdown point as least squares and does not have the bounded influence because it does not take into account the leverage of the observations to down hill in its equations(Hampel *et al.*, 1986).

A generalization of M-regression estimator is given by the generalized M-estimator (GM-estimator). They were suggested in order to maintain efficiency of M-estimator, and to limit the influence of the leverage point simultaneously(Hampel *et al.*, 1986). But Maronna, Busto and Yohai(1979) showed that GM-estimator has breakdown point depending on the dimension of independent variables and it is at most $1/(p+1)$ where p is the number of independent variables in the regression model, including the intercept if present.

Rousseeuw(1984) introduced the least median of squares (LMS) and the least trimmed squares (LTS). These estimators minimize the median and the trimmed mean of the squared residuals respectively. They are first suggested as high breakdown point estimation with regression equivariance. As following, S-estimator(Rousseeuw and Leroy 1987) was suggested as robust estimator with high breakdown point and more efficiency than other high breakdown estimators. MM-estimator(Yohai,1987) and one-step GM estimator(Simpson *et al.*,1992; Coakley and Hettamanspergerare,1993) are multistage estimators with desirable properties, efficiency, high breakdown point and bounded influence while other high breakdown point estimators do not satisfy robustness at the same time. For this purpose, MM-estimator and one-step GM-estimator suggested for simultaneous satisfaction of all or some robustness properties, like efficiency, high breakdown point and bounded influence.

Especially, these multistage estimators have a common properties that they use a high break-

down point estimator, such as LMS, LTS, or S-estimators, as their initial estimates for retaining high breakdown point. But high breakdown point estimation have some drawbacks. First of all, computing any of these estimators exactly is impractical in all but small datasets, because they involve the combinatorial problem of determining how many cases are used. Therefore, they are based on resampling techniques and their solutions are determined randomly(Rousseeuw and Leroy, 1987), and then they can be even inconsistent(Hawkins and Olive, 2002). Second problem is their lower convergence rate. For example, LMS has the low convergence rate as $n^{-1/3}$. It makes direct effect of efficiency of estimates, and, moreover, multistage estimators are not free from it.(He and Portnoy,1992) Third, they do not have bounded influence for $\mathbf{X}$-space or high leverage points although they limitedly affected from response outlying observations.

We introduce a new class of robust regression estimators, self-tuning estimator(STE). It basically inherits spirits of S-estimator and GM-estimator. Moreover, its goal is to construct the estimation method not to depend on data size without loss of appropriate robustness. In Section 2, we suggest computing algorithm for self-tuning regression estimator and introduce some properties revealed until now. In Section 3, we refer to the result of investigation through Monte Carlo study meanwhile and simply allude to further research.

2 Computing algorithm

Consider the multiple regression model,

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \epsilon_j, \quad j = 1, 2, \cdots, n. \tag{1}$$

Let $\bar{y}_{U_i}$ and $\bar{y}_{L_i}$ be the sample means of y_j's based on respective observations with $x_{ij} \geq \bar{x}_i$ and $x_{ij} < \bar{x}_i$ for each $i = 1, 2, \cdots, p$ where $\bar{x}_i$ is the sample mean of the ith independent variable.

step 1 For the ith independent variable, partition n observations into four sets;

$$\begin{aligned}\{(\mathbf{x}_j, y_j), j = 1, 2, \cdots, n\} = &\{(\mathbf{x}_j, y_j), x_{ij} \geq \bar{x}_i \text{ and } y_j \geq \bar{y}_{u_i}\} \\ &\cup \{(\mathbf{x}_j, y_j), x_{ij} \geq \bar{x}_i \text{ and } y_j < \bar{y}_{u_i}\} \\ &\cup \{(\mathbf{x}_j, y_j), x_{ij} < \bar{x}_i \text{ and } y_j \geq \bar{y}_{u_i}\} \\ &\cup \{(\mathbf{x}_j, y_j), x_{ij} < \bar{x}_i \text{ and } y_j < \bar{y}_{u_i}\}\end{aligned}$$

Let $\mathbb{O} = \cup_{i=1}^{p} O_i$ where O_i is the all possible non-empty subsets of the four partitions by the ith independent variable and denote K be the number of sets in $\mathbb{O}$ where each set includes

at least $p+1$ observations to have the OLS for $\beta = \{\beta_0, \beta_1, \cdots, \beta_p\}$. We denote these K sets of observations by $E_1, E_2, \cdots, E_K$.

step 2 Obtain OLS estimate $\mathbf{b}_k^0$ from observations in E_k and apply $\mathbf{b}_k^0$ to n observations to calculate standardized residuals $\tilde{r}_j(\mathbf{b}_k^0) = \frac{r_j(\mathbf{b}_k^0)}{s(\mathbf{b}_k^0)}$ where $r_j(\mathbf{b}_k^0) = y_j - \mathbf{x}_j'\mathbf{b}_k^0$, $j = 1, 2, \cdots, n$ and $s(\mathbf{b}_k^0) = (0.6745)^{-1}\text{median}(|r_j(\mathbf{b}_k^0)|)$. Then we calculate

$$\mathbf{b}^1 = \arg\min \sum_{|r_j(\mathbf{b}_k^0)|<c_1, |r_i(\mathbf{b}_k^0)|<c_1} \left[|r_i(\mathbf{b}_k^0)| - |r_j(\mathbf{b}_k^0)|\right]_+ \tag{2}$$

where c is a cut-off value, $k = 1, 2, \cdots, K$ and $[x]_+ = \max(0, x)$.

step 3 Calculate the preliminary STE $\mathbf{b}_{pSTE}$ from observations satisfying $|\tilde{r}_j(\mathbf{b}^1)| < c$ for $j = 1, 2, \cdots, n$, using $\mathbf{b}^1$ as the initial estimate.

step 4 Remove the observations with $|\tilde{r}_j(\mathbf{b}_{pSTE})| > c_1$ for $j = 1, 2, \cdots, n$. We recall these observations temporary outliers.

step 5 Repeat step1 - step4 until no additional outlier is detected using the remaining observations from step4 of the previous repetition. Accordingly, the sample size n used in step1 - step4 is adjusted in each repetition.

The estimate from step5 is our self-tuning estimate denoted by $\mathbf{b}_{STE}$, applying $\mathbf{b}_{STE}$ whole observations to detect the final outliers such that $|\tilde{r}_j(\mathbf{b}_{STE})| > c_2$ $(c_1 > c_2)$.

There are two main reasons for step1. First reason is that, by partitioning observations by the sample mean of independent variable, and then the sample mean of dependent variable, we can isolate bad leverage points. Because our self-tuning estimate uses OLS estimates as an initial estimate, the STE may depends on outliers since the OLS as an initial estimate depends on outliers. (This is one of weak points of our previous STE which were indicated by the referees for our previous paper. This is actually the concept of high breakdown estimation). Thus, we need to an OLS estimate which is less affected on outliers. Step2 is for this purpose. The second reason is as follows with step2. When an E_k contains outlier or outliers, the OLS fitted on E_k should be distorted and thus most of the resulting residuals calculated from all n observations should be large . Therefore, the corresponding $s(\mathbf{b}_k^0) = (0.6745)^{-1}\text{median}(|r_j(\mathbf{b}_k^0)|)$ is large, and hence $\sum_{|r_j(\mathbf{b}_k^0)|<c_1, |r_i(\mathbf{b}_k^0)|<c_1} \left[|r_i(\mathbf{b}_k^0)| - |r_j(\mathbf{b}_k^0)|\right]_+$ should be large. This means that the OLS on E_k has low possibility to be an initial estimate. Although, from step2, we have an OLS which is less affected on outliers, the OLS may be affected on outliers when they are scattered moderately far from the majority of y over all range of x's because the partitions in step1 operate well for

outliers which are far from the majority of x and y. Thus, we use step3. When outliers reside in symmetric positions of the true linear line, step3 may work for outliers which are farther from the true line. Thus we need repetition to detect outliers in the opposite position. Step4 is for this. Step1 - step4 can be the procedure for data cleaning and for obtaining an initial estimate not depending on outliers for our final STE in step5. In our simulation study, we use $c_1 = 4$ and $c_2 = 3$.

3 Conclusion

Out purpose is to construct to estimation method not to depend on data size with appropriate robustness simultaneously. Monte Carlo study shows that STE has high breakdown point, unbounded influence and high efficiency. Moreover, STE always guarantees unique solution and consume less computing time because it does not depend on data size unlike other high breakdown point estimators, and it does not need resampling technique for estimation.

We do not make a close and enough examination for STE yet. For example, its asymptotic properties and robustness are not justified and do not go into details in theoretical sense.

References

Coakley,C.W. and Hettamansperger,T.P. (1993). A bounded influence, high breakdown, efficiency regression estimator. *Journal of the American Statistical Association*, **88**, 872-880.

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

Hampel,F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383-393.

Hampel,F.R., Ronchetti,E.Z., Rousseeuw,P.J. and Stahel,W.A. (1986). *Robust Statistics. The approach based on influence functions.* Wiley, New York.

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*, **79**, 871-880.

He, X. and Portnoy, S. (1992). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, **20**, 2161-2167.

Huber,P.J. (1973). Robust regression : asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799-821.

Huber,P.J. (1981). *Robust Statistics.* Wiley, New York.

Maronna,R.A., Butos,O.H. and Yohai,V.J. (1979). Bias and efficiency robustness of general M-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation* (Gasser,T. and Rocenblatt,M., eds.) 91-116. Springer-Verlag, New York.

Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871-880.

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

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.

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