Gazi University Journal of Science
GU J Sci
24(4):747-752 (2011)

GUJS

The Comparing of S-estimator and M-estimators in Linear Regression

Meral ÇETİN[1], Onur TOKA[1♠]

[1]*Department of Statistics, Faculty of Science, Hacettepe University, 06800, Ankara, Turkey*

Received: 11.10.2010 Revised: 23.11.2010 Accepted:21.12.10

ABSTRACT

In the presence of outliers, least squares estimation is inefficient and can be biased. In the 1980's several alternatives to M-estimation were proposed as attempts to overcome the lack of resistance. Least Trimmed Squares (LTS) is a viable alternative and is presently the preferred choice of Rousseeuw and Ryan (1997, 2008). Another proposed solution was S-estimation. This method finds a line that minimizes a robust estimate of the scale of the residuals. This method is highly resistant to leverage points, and is robust to outliers in the response. However, this method was also found to be inefficient.

The aim of this study is to compare S-estimator with other robust estimators and the least squares estimators and also an example is given to illustrate the efficiency of S-estimator. The data used in this example are the air pollution measures. And finally a simulation study has been presented in this study.

Key words: *M-estimators, S-estimator, Robust regression, Least median squares ,Air pollution*

1. INTRODUCTION

It is well known that least squares estimator in the linear regression model is not robust: only one observation can give rise to arbitrarily poor estimates of the coefficients.

Hampel [7] proposed an estimator based on minimizing the median absolute deviation of the residuals and gave its breakdown point as 1/2, the highest possible value for affine equvariant estimates. S estimator is the generalization of least median squares.

Rousseeuw and Yohai[13] introduced S-estimator in framework of multiple regression. They proposed to call them S-estimators because they are based on estimators of scale. These estimators were shown to have the same asymptotic properties as M estimators. Also these estimators have good robustness proporties, as their breakdown point has 50%.

Davies [3] investigated some properties of S-estimators of multivariate location and covariance. He studied consistency, asymptotic normality and breakdown point using different definition from the one given by Rousseeuw and Yohai[13].

2. STATISTICAL BACKROUND AND METHODS

The general linear regression model is given $y_i = x_i^t \beta + \varepsilon_i$ for i=1,...,n.

Where x_i and β are p-dimensional column vectors, ε_i is the error term with $E(\varepsilon_i)=0$ and $Var(\varepsilon_i)=\sigma^2$. Our aim is to estimate the unknown regression parameter β. The most popular estimation technique is the classical least squares method and can be defined as following, where e is the residual term,

$$\text{minimize} \sum_{i=1}^{n} (y_i - x_i^t b)^2 = \sum_{i=1}^{n} e_i^2 \qquad (2.1)$$

However, this estimator is not robust, because the occurance of even one outlier can spoil the estimates very badly.

In connection with this effect, the breakdown point of the least squares estimator is 1/n and approaches zero when n tends to infinity([4]). The breakdown point was

♠Corresponding author, e-mail: onur.toka@hacettepe.edu.tr

introduced by Hampel [6]. In words, the breakdown point is the smallest fraction of contaminated data.

To find more robust estimators, first step came from Edgeworth [5]. His least absolute values or L_1 criterion is,

$$\text{minimize} \sum_{i=1}^{n} \left| \; y_i - x_i^t b \; \right| \qquad (2.2)$$

But this estimator still cannot cope with outlying x_i and breakdown point=1/n $\rightarrow$0.

Huber [8,9] proposed the regression M-estimators as an alternative robust regression estimator to the least squares. This method based on the idea of replacing e_i^2 in equation (2.1) by $\rho(e_i)$. An M-estimate minimizes the following function

$$\text{minimize} \sum_{i=1}^{n} \rho(e_i) \qquad (2.3)$$

where $\rho(e_i)$ is a symmetric function with a unique minimum at zero. A widely used $\rho(e_i)$ function is the Huber function and M-estimators obtained from this function are sometimes called Huber M-estimates or monotone M-estimates. Huber M-estimator is robust against outliers in y-direction, but it is not robust against outliers in x-direction. Because of vulnerability to leverage points, generalized M-estimators (GM-estimator, for short) were considered ([10]). The basic idea is to bound influence of outlying x_i using some weight function. GM estimator for regression do not have good global robustness properties measured by the breakdown point which can not be greater than 1/(p+1). This problem has been solved by the one-step GM-estimator that have high breakdown and bounded influence function.

All this raises the question whether it is possible to obtained robust regression with high breakdown point. The first affirmative answer was given by Siegel [14] and he proposed the repeated median (RM). This estimator has 50% breakdown point but is not affine equivariant, it depends on the choice of the coordinate axes of the x_i. The next high breakdown point estimator was the least median of squares (LMS**)** estimator ([11]), has 50% breakdown point and is a affine equivariant. But it does not have good asymptotic properties. The other estimator with high breakdown point is the least trimmed squares (LTS) estimator. It is obtained from

$$\text{minimize} \sum_{i=1}^{h} e_{(i)}^2$$

where $e_{(1)}^2, e_{(2)}^2, \ldots, e_{(n)}^2$ are the ordered squared residuals, from smallest to largest, and the value of h must be determined. Rousseeuw proposed LTS in a symposium in 1983.

A good source is Rousseeuw[11].

Rousseeuw and Yohai[13] introduced S-estimators which is the generalization of least median squares. S-estimators are last HBP estimators. They proposed to call them S-estimators because they are based on estimators of scale. These estimators were shown to have the same asymptotic properties as M estimators of regression.

2.1 S Estimator

A generalization of least median squares was given by Rousseeuw and Yohai[13] who introduced a new class of estimator, S-estimator. They proved consistency and asymptotic normality for a restricted class of S-estimator.

Let ρ be a symmetric, continuously differentiable function such that ρ (0)=0 and is strictly increasing on [0,c]. Let k= $\int \rho(x) d\Phi(x)$, where Φ is the standard normal distribution.

Rousseeuw and Yohai[13] introduced so-called S-estimator, which is derived from a scale statistics in an implicit way, corresponding to s(θ)

minimize s(θ),

where s(θ) is a certain type of robust M-estimate of the scale of the residuals $e_i(\theta),\ldots,e_n(\theta)$.

S-estimator ([13]) constitutes another class of high breakdown affine equivariant estimators with convergence rate $n^{-1/2}$. They are defined by minimization of the dispertion of the residuals:

$$\underset{\hat{\theta}}{\text{minimize}}\; s(e_1(\hat{\theta}),\ldots,e_n(\hat{\theta})) \qquad (2.4)$$

with final scale estimate

$$\hat{\sigma} = s(e_1(\hat{\theta}),\ldots,e_n(\hat{\theta}))$$

The dispersion $s(e_1(\theta),\ldots,e_n(\theta))$ is defined as the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = K. \qquad (2.5)$$

K is often put equal to $E_\Phi[\rho]$, where Φ is the standard normal. The function ρ must satisfy the following conditions:

(1) ρ is symmetric and continuously differantiable, and ρ (0)=0.

(2) There exists c>0 such that ρ is strictly increasing on [0,c] and constant on [c, ∞).

If there are more than one solution from (2.5), then $s(e_1,\ldots e_n)$ equal to the supremum of the set of solutions ([12,13]).

In equation (2.5), ρ is taken Tukey's biweight function and hyperbolic tangent estimator. Tukey's biweight function is given by

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{for } |x| \le c \\ \frac{c^2}{6} & \text{for } |x| > c \end{cases} \quad (2.6)$$

ρ will depend on a positive tuning parameter c as $\rho_c(t) = c^2\rho(t/c)$. The turning parameter plays a very important role for the asymptotic and breakdown properties of s-estimator for regression. For all values of c, $\rho(\infty)$ will be the same so that to obtain ½ breakdown point. To get an S-estimator with the asymptotic breakdown points, one can choose c=2.937, 3.42 and 4.00 respectively. Setting c=1.5476 gives 1/2 breakdown point but the asymptotic relative efficiency (ARE) of the estimator for b is 28.7%. On the other hand the ARE is 91.7% for c=4.096, but breakdown point is 15%. This demonsrated that one has trade-off between high breakdown point and high asymptotic relative efficiency ([1])

3. APPLICATIONS

The aim of this study is to compare S-estimator with other robust estimators and the least squares estimators and also an example is given to illustrate the efficiency of S-estimator. The data used in this example are the air pollution measures between 1995:01-1995:04.([2]).

The variables are,

y: sulphur dioxide (SO2)

x1: pressure

x2: minimum daily temperature,

x3:humidity

x4: sun

x5: rain

x6: speed of wind

x7: direction of win

G: dummy variable

Months are plugged into the model as a dummy variable and multivariate general model is shown as follows, where e is the residual term.

$$SO_2 = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8G + e,$$

In this model, the weak collinearity is detected among the explanatory variables and the outlier is not present after the examination. The parameter estimations of the model are determined based on LS, Huber, LTS and S-estimator by using S-Plus.

4. RESULTS AND DISCUSSION

The results of the parameter estimations are given Table 1. The estimates of the parameters are very close to the actual, values except constant term. The difference in the constant term may be because of collinearity issue.

Table 1. The results of the parameter estimations

	LS	Huber	LTS	S
Constant	0.9862	1.7849	2.5801	0.7626
b_1	0.0047	0.0039	0.0032	0.0050
b_2	-0.0336	-0.0352	-0.0387	-0.0355
b_3	-0.0014	-0.0015	-0.0018	-0.0016
b_4	-0.0003	-0.0004	-0.0003	-0.0005
b_5	-0.0378	-0.0389	-0.0418	-0.0410
b_6	0.0093	0.0132	0.0217	0.0190
b_7	-0.0398	-0.0420	-0.0519	-0.0414
b_8	0.3021	0.2855	0.2054	0.2718

There is no outlier in the data set. However, we try to find out the effects of outliers in the estimated parameters and compare estimation of the parameters by replacing an outlier with a datum in response variable. And the same procedure is repeated for the case of two outliers cases in response variable. The results of the parameter estimations of the model for one outlier and two outliers cases are presented in Tables 2-3.

Table 2. The results of the parameter estimations for one outlier.

	LS	Huber	LTS	S
Constant	9.5637	4.3804	7.2808	2.4112
b_1	-0.0036	0.0014	-0.0016	0.0034
b_2	-0.0336	-0.035	-0.0348	-0.0355
b_3	-0.0018	-0.0019	0.00026	-0.00206
b_4	0.0005	-0.0002	-0.00037	-0.00043
b_5	-0.0296	-0.0367	-0.0436	-0.0395
b_6	-0.0037	0.0087	0.0158	0.01614
b_7	-0.0577	-0.0483	-0.0388	-0.0457
b_8	0.3342	0.2983	0.2346	0.2834

Table 3. The results of the parameter estimations for two outlier.

	LS	Huber	LTS	S
Constant	11.2018	4.7731	1.6591	2.3818
b_1	-0.0058	0.00081	0.0039	0.0033
b_2	-0.0281	-0.0328	-0.0329	-0.0342
b_3	0.00198	0.0000106	0.00019	-0.0012
b_4	0.00115	0.0000418	-0.00051	-0.00030
b_5	-0.0257	-0.03525	-0.03021	-0.03896
b_6	0.0043	0.01256	0.01713	0.01821
b_7	-0.0401	-0.0401	-0.03582	-0.0417
b_8	0.3094	0.2809	0.2466	0.2746

The estimates of the parameter differ in Table 2 when they are compared with the values in Table 1. The LS estimation is affected by the outlier, as excepted. While Huber and S-estimator stay unchanged, LTS estimator is affected a little. However S-estimator in terms of values of parameter estimates is approximately the same as those in Table 1.

The results of the parameter estimations of the model for two outliers case are presented in Table 3. It can be seen that the Huber and LTS estimators are less affected than the LS estimator, When there are two outliers, S-estimator stays unchanged and has similar results as in Tables 1-2.

As a result, it can be infered that the number and magnitude of outlier do not affect S-estimator.

4.1 Simulation Study

In this paper, in order to compare robust estimators with LS, a simulation study has been presented. In order to obtain the MSE of the estimators, a program was coded by using S-Plus functions. We generated 15 independent replicates of five independent uniform random variables on U[-1,+1] and 15 independent normally distributed errors e_i with expectation 0 and variance σ^2 =0.01, 1, 10 and 100 respectively. Then we generated observations for y_i according to the following model:

Beta1: ($5,3,\sqrt{6},0,0$)

It is designed to consider model without intercept**.** In order to see the effects of the outliers on the estimators, the estimators are examinated in cases of no outlier, one outlier and two outliers for e. Also, the MSE values of the estimators according to σ^2 are given in Tables 4-5.

Table 4. MSE of estimators in case of no outlier

	σ^2 =100	σ^2 =10	σ^2 =1	σ^2 =0.01
LS	0.062710	0.0671099	0.06271099	0.06271099
Huber	0.067337	0.0673375	0.06733756	0.06733756
LTS	0.129084	0.1288023	0.12943370	0.1282498
S	0.134558	-	-	-

Table 5. MSE of estimators in case of one outliers

	σ^2 =100	σ^2 =10	σ^2 =1	σ^2 =0.01
LS	0.09231048	0.3716438	3.156176	309.0605
Huber	0.07472932	0.07467774	0.07472412	0.0761199
LTS	0.11276860	0.1072625	0.1080733	0.1102873
S	0.1315891	-	-	-

The hyphen in Tables 4-6 means that we have no result as the rank is equal to 0.

Table 6. MSE of estimators in case of two outliers

	σ^2 =100	σ^2 =10	σ^2 =1	σ^2 =0.01
LS	0.1183422	0.6705631	6.183873	612.2494
Huber	0.09996891	0.1037333	0.1049549	0.1252048
ITS	0.1500158	0.09320952	0.0945125	0.09408558
S	0.2261035	-	-	-

According to the Tables, the MSE of estimators are similar for the entire variance situations in the case of no outliers. Naturally, the most effected estimator from outliers is LS. When the variance is 100, S estimator is not affected from outliers. Equally Huber and LTS estimators are not affected from outliers for all values of variance. Moreover, it is seen that LTS has the smallest MSE even if there are two outliers.

Generally, as expected, robust estimators have better performance than LS in case of outliers.

REFERENCES

[1] Arslan O., O. Edlund, H. Ekblom, "Algorithms to compute CM- and S-estimates for regression", ***Metrika***, 55, 37-51 (2002).

[2] Candan, M., "Robust Estimators in linear Regression Analysis", MSc. Thesis, ***Hacettepe University, Department of Statistics***, 94 (1995).

[3] Davies, P. L., "Asymptotic Behavior of S-estimates of multivariate location parameters and dispersion matrices", Technical Report, ***University of Essen***, West-Germany, (1987).

[4] Dohono, D.L. and Huber, P. J., "The notion of breakdown point, in A Festschrift for Erich Lehmann", edited by P. Bickel, K. Doksum, and J. L. Hodges, ***Jr., Wadsworth***, Belmont, CA., (1983).

[5] Edgeworth, F. Y., "On observations relating to several quantities", ***Hermathena,*** 6: 279-285, (1887).

[6] Hampel, F. R., "A general qualitative definition of robustness", ***Ann. Math. Stat***., 42: 1887-1896, (1971).

[7] Hampel, F.R., "Beyond location parameters: Robust concepts and methods", ***Bull. Int. Stat. Inst.,*** 46: 375-382, (1975).

[8] Huber, P. J., "Robust regression: Asymptotics, conjectures and Monte Carlo", ***Ann. Stat.,*** 1:799-821, (1973).

[9] Huber, P. J., "Robust Statistics", ***John Wiley & Sons,*** New York, (1981).

[10] Moranno, R. A., Bustos, O., and Yohai, V., "Bias-and efficiency-robustness of general M-estimators for regression with random carriers, in Smoothing Techniques for Curve Estimation", edited by T. Gasser and M.Rosenblatt, ***Spinger Verlag***, New York, 91-116, (1979).

[11] Rousseeuw, P. J., "Least median of squares regression", ***J. Am. Stat. Assoc***., 79: 871-880, (1984).

[12] .Rousseeuw P. J and Leroy, A. M., "Robust regression and outlier detection", ***Wiley,*** New York, (1987).

[13] Rousseeuw, P.J., and Yohai, V., "Robust regression by means of S-estimators, in Robust and Nonlinear Time Series Analysis", edited by J. Franke, W. Hardle, and R.D. Martin, Lecture Notes in Statistics No.26, ***Spinger Verlag***, New York, 256-272, (1984).

[14] Siegel,A. F., "Robust regression using repeated medians", ***Biometrika***, 69: 242-24.