行政院國家科學委員會專題研究計畫成果報告

遺漏資料的穩健有效迴歸估計

Very Robust and Efficient Regression in the Presence of Missing Values

計畫編號：NSC 89－2118－M－004－007
執行期限：88 年 12 月 1 日至 89 年 7 月 31 日
主持人：鄭宗記　執行機構及單位名稱：國立政治大學統計系

中文摘要

關於判斷偵測資料中可能存在的許多個離群質的問題，基本上需要「高破壞點估計式」(high breakdown estimator)。Atkinson 和 Cheng (2000)提出一方法處理資料中含有遺漏值時的線性迴歸模型的離群值偵測問題。基本上，其利用 Atkison (1994)的「前進搜尋演算法」(forward search algorithm)及結合了 EM 演算法 (Dempster *et al.* 1977)。大多數的高破壞點估計式伴隨著較低的有效性(efficiency)，而遺漏資料又將影響到估計的有效性，因此，本研究探討、檢驗一些高破壞點估計式在有限樣本及不完全資料的狀況下的一些特性。首先推廣 Atkinson 和 Cheng(1999)之結果至其他高破壞點估計式，然後藉由電腦模擬結果，針對資料中離群值、遺漏值所佔之比例，參數維度及樣本大小等因素來比較這些穩健估計式之有效性及其對離群值偵測的結果。

關鍵詞：EM 演算法、離群值偵測、前進搜尋演算法、高破壞點估計式、遺漏值

Abstract

To deal with the problem of detection of multiple outliers, a high breakdown estimator is essentially required. Atkinson and Cheng (2000) propose a method that uses the forward search algorithm for robust regression problem to extend missing value techniques to data with several outliers. However, most high breakdown estimators are accompanied by low efficiency. Yohai (1987), Yohai and Zamar (1988) and Rousseeuw and Croux (1994) proposed some high breakdown estimators with high efficiency. Cheng (1998) conducts a simulation study to present the finite-sample efficiency of some high breakdown estimators, such as least median of squares (LMS), least trimmed squares (LTS) and least quartile difference (LQD) estimators. It is of interest to examine the performance and efficiency of these three estimators when part of the data is missing. Some real data examples are used to illustrate and compare the resulting algorithms. Simulation experiment will be carried out to show the results..

Keywords: EM algorithm, detection of multiple outliers, forward search algorithm, high breakdown estimator, missing values

1. Introduction

The detection of multiple outliers has been a particularly intractable problem, which essentially requires a high breakdown estimator. However, most high breakdown estimators are accompanied by low efficiency. Some high breakdown estimators retaining a higher efficiency have been proposed (Yohai 1987, Yohai and Zamar 1988, and Rousseeuw and Croux 1994). Stefanski (1991) discussed the conflict between efficiency and high breakdown point. Morgenthaler (1991) proved that positive breakdown estimators may have low finite sample efficiency relative to least squares estimators. On the other hand, several authors argue that the concept of ``efficiency at the model'' is not a robust concept (see Davies

(1994)). In the study of linear regression, Davies (1993) investigated several desirable aspects of robustness, which however do not include efficiency. He also argues that for linear regression with normal errors there is a tradeoff between gross error sensitivity and efficiency, and between breakdown point and efficiency. The discussion of Atkinson and Cheng (1999) shows that the least trimmed squares (LTS) estimates provide very robust methods to identify multiple outliers as well as parameter estimates with a high efficiency than those done by the least median of squares (LMS) for the linear regression model. The efficiency of the LTS is related to the quantile index, q. The higher the value of the quantile index, the higher the efficiency.

The asymptotic result of Hössjer *et al.* (1994) shows that the efficiency of the LMS approximates to zero as the sample size goes to infinity, and that of the *S*-estimators is at most 33%. Coakley (1991) also shows that the efficiency of the LTS surpasses 1/3, 1/2 and 3/4 for $q<0.8n$, $0.88n$ and $0.96n$, respectively. With the same high breakdown for an *S*-estimator, a generalised *S*-estimator proposed by Croux *et al.* (1994) can have asymptotic efficiency 68.4%. This implies that using almost 90% of the data to fit the LTS still has lower asymptotic efficiency than the LQD, but for this case, the former one only has breakdown 10%. In simple linear regression, the finite-sample efficiencies of LTS and LMS are actually higher than their asymptotic efficiencies under Gaussian errors (Coakley *et al.* 1994). You (1999) also compare the finite sample performance of some high breakdown estimators, which is carried out by a Monte Carlo study. The simulation results of Atkinson and Cheng (1999) also show the same tendency for the multiple regression model. Therefore, it is of interest to inspect the finite-sample efficiencies of the LMS, LTS and LQD, especially in high dimensional problems.

Often a part of the data is missing. The current approaches dealing with missing values also inherit the same problems of outliers as arise in complete data. Both the EM algorithm (Dempster et al. 1977) and multiple imputation (Rubin 1987) rely on the assumption of normality, and hence tend to be sensitive to outliers. As a result, they may fill in abnormal values to missing variables or tend to move (potential) outlying observations to the center of the data if one ignores the existing outliers. Therefore the masking and swamping effects due to multiple outliers appear even more serious if some data are missing (Atkinson and Cheng 2000). In this paper, we also extend the work of Atkinson and Cheng (2000) to LQD for incomplete data. The algorithms lead to the identification of multiple outliers from incomplete data. Because of the properties of the forward search algorithm, we can simultaneously obtain more reasonable imputed values for missing values as well as very robust parameter estimates.

Several algorithms for high breakdown estimators have been proposed in the last decade. Computational time is almost no longer a particularly difficult issue in the field of robust statistics (Rousseeuw and Van Driessen 1999 a and b). Among the existing algorithms, the forward search algorithm proposed by Atkinson (1994) is a very fast and efficient one. It has been first used to find the LMS and then extended to LTS by Atkinson and Cheng (1999). A framework of this algorithm and its other extensions have been described in Atkinson and Riani (2000). One of the purposes of robust statistics is the detection of outliers. Furthermore, it may not need an exact solution for the purpose of outlier detection (Atkinson 1994). Hence, it is also of particular interest to verify the problem of detection of outliers using high breakdown estimators with low or high efficiency. The relationship among the high breakdown point, efficiency and detection of outliers remains vague. In this paper, a simulation experiment is conducted to clarify the problem. Three high breakdown point estimators, LMS, LTS and LQD, are used to compare their finite sample efficiencies and the abilities to identify outliers for both complete and incomplete data.

2. The forward search algorithm for the LQD

Because of the advantage of the forward search algorithm, we also apply it to the search of the LQD. The algorithm for the LQD is similar to those for the LMS and LTS. The details of the algorithm can be found in Atkinson (1994) and Atkinson and Cheng (1999). The only difference is that we intend to find the scale estimate

$$\tilde{\sigma} = |e_i - e_j|_{\binom{h_p}{2}\binom{n}{2}}$$

where $\tilde{\sigma}$ is $\binom{h_p}{2}$th order statistic among the $\binom{n}{2}$ elements of the set $\{|e_i - e_j|; i < j\}$. It is used to Studentise the residuals and evaluate the performance of each search. The scale estimate can be obtained by the time-efficient algorithm of Croux and Rousseeuw (1992). Therefore, the algorithm has the same computation time as that for the LMS.

For the problem of missing values, we extend the approach of Atkinson and Cheng (2000) to the LQD as well, which combines the forward search algorithm with EM to deal with outliers in missing data.

3. Simulation results

In this section, we compare the finite-sample efficiencies of three estimators, LMS, LTS and LQD, as well examining the ability to detect outliers via a Monte Carlo study. To implement the simulation experiment, the partition algorithm which was proposed by Woodruff and Rocke (1994) for multivariate data is also used to obtain the initial subset for the purpose of saving computation time.

3.1 Complete data

The simulation experiment is quite similar to that in Atkinson and Cheng (1999). 100 replications are generated to compute the variance of estimated parameters and compare the detection of outliers.

To simplify the simulation process, we consider a fixed number of searches instead of the factor of time for the partition algorithm, in which each cell contains at least *1.5p* cases, normally *2 p*. In each cell, 500 elemental subsets are randomly chosen for combinatorial search. For the sake of computational time, we first generate a data set containing no outliers and use 100 forward searches to get the Studentised residuals. This means that we apply the old procedure to the standardisation, but the same $\boldsymbol{X}$ matrix carries over all 100 simulation runs.

Table 1 presents the simulations results of the LMS, LTS and LQD based on the factors of dimension, sample size and proportion of outliers. The value q of the LTS is chosen as [n (1-(proportion of outliers) -10%)]. The following quantity is used to evaluate the efficiencies of the three estimators,

$$\frac{\sum_{i=0}^{p-1} Var(\hat{\beta}_i)}{p} \times 10^4 \qquad (1)$$

where $Var(\hat{\beta}_i)$ is calculated from 100 simulation runs. Overall, the LMS is inferior to both LTS and LQD in terms of (1). For each row of Table 1 except the last three ones, we can see that the values (1) of all three estimators decreases when the sample size increases whatever the dimension and proportion of outliers, which means that the efficiency increases as the sample size increases. The rate of decrease of (1) for LQD is faster than for the other two. For a fixed sample size, the LTS can have higher efficiency than the LQD when the proportion of outliers is less than 20%, but the latter is better when dimension is higher (e.g. 20) and/or sample size is large (e.g. 200).

The LMS and LTS actually have higher finite-sample efficiency than indicated by the asymptotic results. The LQD is the best when both dimension and sample size are higher. The ability to identify existing outliers is quite similar for these three estimators in our limited simulation. However, the masking effect due to the LMS is more serious than other two. The LTS gives very stable results of detection of outliers whatever the sample size is. When sample size is small the LQD has the

problem of local instability as the LMS does. This problem may disappear when the sample size is large for the LQD, but not for the LMS. A remark is that we can increase the probability of success by increasing the number of searches of both combinatorial search and forward search algorithm.

Table 1 Simulation results of the quantity (1) from 100 simulation runs using the LMS, LTS and LQD.

Dimension (p)	Proportion of bad data	Estimates	Sample size (n) 50	100	200	400
5	10%	LMS	76.1708	52.2976	27.1206	12.3214
		LTS	49.9384	20.4625	11.2884	5.0065
		LQD	62.0816	29.0419	10.8513	4.3453
	20%	LMS	71.4396	32.4844	22.7636	15.8314
		LTS	49.4509	24.5567	13.8512	7.3217
		LQD	54.4571	24.0940	12.6020	5.9636
	30%	LMS	59.2898	36.8278	20.0111	9.3316
		LTS	62.3270	37.5063	18.3479	7.8831
		LQD	54.3693	31.8617	14.1124	5.6362
10	10%	LMS	115.4784	73.8874	26.7364	17.6271
		LTS	62.7102	36.1857	10.1563	6.2680
		LQD	106.0457	41.3584	10.4333	6.8414
	20%	LMS	92.1499	41.3976	30.8812	13.1560
		LTS	86.1053	31.8233	18.8175	8.6295
		LQD	73.8047	27.0876	16.0780	6.6306
	30%	LMS	72.1808	33.2419	21.1842	10.4943
		LTS	103.0409	32.6180	17.8146	9.3144
		LQD	63.7669	25.3397	12.4909	8.4275
20	10%	LMS	--	58.7338	31.1014	17.3836
		LTS	--	30.4690	15.1072	6.3709
		LQD	--	50.7163	19.4341	7.2644
	20%	LMS	--	51.9438	25.3576	10.4232
		LTS	--	46.9662	99%	8.1283
		LQD	--	38.6709	14.6601	6.1380
	30%	LMS	--	90%	30%	76%
		LTS	--	88%	31%	66%
		LQD	--	88%	31%	62%
30	10%	LMS	--	--	23.7825	14.5917
		LTS	--	--	16.1701	7.7262
		LQD	--	--	19.4184	9.2235
	20%	LMS	--	--	25.1213	11.7316
		LTS	--	--	17.3618	8.0131
		LQD	--	--	18.7156	6.6621
	30%	LMS	--	--	90%	10.3175
		LTS	--	--	80%	11.4150
		LQD	--	--	95%	6.4653

3.2 Incomplete data

The similar simulation design is also used for the incomplete data problems. Because of the slow convergence of the EM algorithm, the simulation scale has been reduced. Here only present part of the

completed results. Figure 1 and 2 show the boxplots of the regression coefficient estimates. There are 5 percent of elements of the design matrix assumed to be missing complete at random. The dimension is 5 and sample sizes are 100, 200 and 400. 10 and 20 percent of data are outliers in Figure 1 and 2, respectively. Both figures show that there is no very significant difference among these three estimates. This is quite different from the previous subsection. We suspect that the results would be improved when the full forward search would be implemented. However, this will increase the computational time very much.

Figure 1 10% outliers and 5% missing in the data

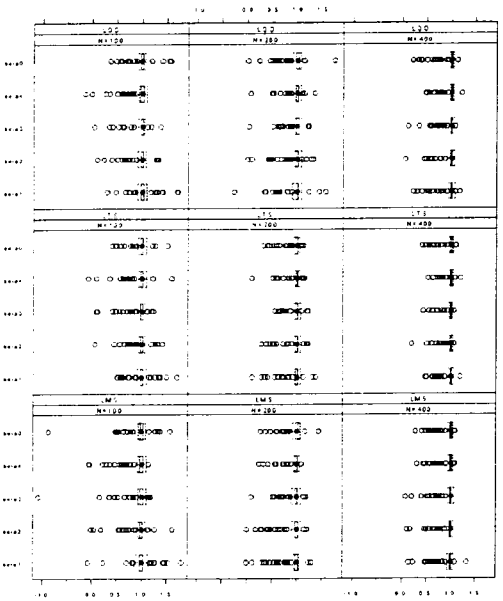
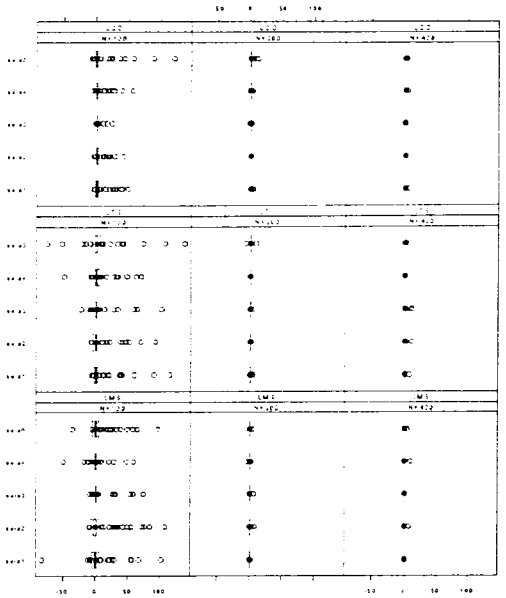


Figure 2 20% outliers and 5% missing in the data



4. Conclusions

We discussed the relationship of efficiency and high breakdown point in this paper. Simulations are carried out with the comparison of the behaviour among three high breakdown point estimators, LMS, LTS and LQD, under finite samples. Results shows that the LMS and LTS actually have higher finite sample efficiency than theory suggests. The efficiency of the LQD appears significantly superior to the other two estimators when the dimension and sample are higher. However it does not guarantee a better performance in the detection of outliers.

Using the time-efficient algorithm of Croux and Rousseeuw (1992), the LQD can only consume the same computation time as the LMS, and it does not need any tuning parameter. The LTS needs more computational time and the choice of quantile index, but it provides very stable results of the identification of outliers no matter what the sample size and dimension are. The LMS has the tendency for identifying inlying observations as outliers. The LQD also has the problem of local instability, but this can be improved when sample size is large. If time is allowed, the

implementation of these three estimators at the same time can provide us a reassured result of the detection of outliers.

From the view of identification of outliers, because of the high breakdown point of the three estimators, they can resist a very high proportion of outliers provided a good algorithm is used for their calculation. For finite samples, they have quite similar performance in terms of the detection of existing outliers. These three estimators can be used as the initial point of some compound estimation procedure. We actually think that it may be quite enough if use of a good algorithm, leading to the identification of outliers, is followed by their rejection and then by fitting by least squares.

The simulation results for the problems of missing values do not look quite satisfied as shown in complete data. It is partly due to the computational issues. Heavy computation remains a problem to apply high breakdown estimators in missing data. Therefore, some further examinations need to be done in the near future to verify the problems discussed in this paper.

5. References

[1] Atkinson, A. C. (1994) ``Fast Very Robust Methods for the Detection of Multiple Outliers", *Journal of the American Statistical Association*, **89**, 1329-1339.

[2] Atkinson, A. C. and Cheng, T.-C. (1999) ``Computing the Least Trimmed Squares Regression with the Forward Search", *Statistics and Computing*, **9**, 251-263.

[3] Atkinson, A. C. and Cheng, T.-C. (2000) ``On Robust Linear Regression with Incomplete Data", *Computational Statistics and Data Analysis*, **33**, 361-380.

[4] Atkinson, A. C. and Riana, M. (2000) *Robust Diagnostic and Regression Analysis*, New York: Springer-Verlag.

[5] Cheng, T.-C. (1998) *Very Robust Statistics in the Presence of Missing Data*, PhD thesis, Department of Statistics, London School of Economics.

[6] Coakley, C. W. (1991) Advances in the Study of Breakdown and Resistance, PhD dissertation. Department of Statistics, Pennsylvania State University.

[7] Coakley, C. W., Mili, L. and Cheniae, M. G. (1994) ``Effect of Leverage on the Finite Sample Efficiencies of High Breakdown Estimators", *Statistics & Probability Letters*, **19**, 399-408.

[8] Croux, C. and Rousseeuw, P. J. (1992) ``Time-Efficient Algorithms for Two Highly Robust Estimators of Scale" in *Computational Statistics*, Vol. 1, eds., Y. Dodge and J. Whittaker, Heidelberg: Physika-Verlag, pp. 441-428.

[9] Croux, C., Rousseeuw, P. J. and H*ö* ssjer, O. (1994) ``Generalized S-estimators", *Journal of the American Statistical Association*, **89**, 1271-1281.

[10] Davies, L. (1994), ``Desirable Properties, breakdown and Efficiency in the Linear Regression Model", *Statistics & Probability Letters*, **19**, 361-370.

[11] Davies, P. L. (1993) ``Aspects of Robust Linear Regression", *The Annals of Statistics*, **21**, 1843-1899.

[12] Dempster, A. P., Laird, M. and Rubin, D. B. (1977) ``Maximum Likelihood from Incomplete Data via the EM algorithm", *Journal of the Royal Statistical Society*, Ser. B, **39**, 1-38.

[13] H*ö* ssjer, O., Croux, C. and Rousseeuw, P. J. (1994) ``Asymptotics of Generalized *S*-estimators", *Journal of Multivariate Analysis*, **51**, 148-177.

[14] Morgenthaler, S. (1991) `` A Note on Efficient Regression Estimators with Positive Breakdown Point", *Statistics & Probability Letters*}, **11,** 469-472.

[15] Rousseeuw, P. J. (1984) ``Least Median of Squares Regression", *Journal of the American Statistical Association*, **79**, 871-880.

[16] Rousseeuw, P. J. and Croux, C. (1993) ``Alternatives to the Median Absolute Deviation", *Journal of the American Statistical Association*, **88**, 1273-1283.

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

[18] Rousseeuw, P.J. and Van Driessen, K. (1999a) ``A Fast Algorithm for the Minimum Covariance Determinant Estimator", *Technometrics*, **41**, 212-223.

[19] Rousseeuw, P.J. and Van Driessen, K.

(1999b) ``Computing LTS Regression for Large Data Sets", Technical Report, University of Antwerp.

[20] Rubin, D. B. (1987) *Using Multiple Imputations to Handle Nonresponse in Sample Surveys*, New York: John Wiley.

[21] Stefanski, L. A. (1991) ``A Note on High Breakdown Point Estimators", *Statistics & Probability Letters*, **11**, 353-358.

[22] Woodruff, D. L. and Rocke, D. M. (1994) ``Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators", *Journal of the American Statistical Association*, **89**, 888-896.

[23] Yohai, V. J. (1987) ``High breakdown point and High Efficiency Robust Estimators for Regression", *The Annals of Statistics*, **15**, 462-656.

[24] Yohai, V. J. and Zamar, R. H. (1988) ``High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale", *Journal of the American Statistical Association*, **83**, 406-413.

[25] You, J. (1999) ``A Monte Carlo Comparison of Several High Breakdown and Efficient Estimators", *Computational Statistics & Data Analysis*, **30**, 205-219.