Proceedings of 11th International Conference on Computer and Information Technology (ICCIT 2008)
25-27 December, 2008, Khulna, Bangladesh

Regression Diagnostics in Large and High Dimensional Data

A. A. M. Nurunnabi[†], and Mohammed Nasser[‡]

[†]School of Business, Uttara University, Dhaka-1230, Bangladesh,
E-mail: pyal1471@yahoo.com
[‡]Department of Statistics, Rajshahi University, Rajshahi-6205, Bangladesh,
E-mail: mnasser.ru@gmail.com

Abstract — **"Learning methods" play a key role in the fields of statistics, data mining, and artificial intelligence, intersecting with areas of engineering and other disciplines. These methods for analyzing and modeling data come in two flavors: supervised and unsupervised learning. Regression analysis and classification are two well known supervised learning techniques. To get an effective model from regression analysis it is necessary to check and preprocess the data set in astronomy, bio-informatics, image analysis, computer vision etc, especially when the data sets are large and high dimensional. In these industries large or fat data appear with unusual observations (outliers) very naturally. Checking raw data for outliers in regression is regression diagnostics. Most of the popular diagnostic methods are not good enough for large and high dimensional data. The aim of this paper is to provide a new measure for identifying influential observations in linear regression for large high dimensional data.**

Index Terms — **Data mining, high dimensional data, influential observation, outlier, regression diagnostics, supervised learning.**

I. Introduction

Data and knowledge mining/discovery is learning from data. Learning means extraction of the regularity from the data. Learning methods for analyzing and modeling data are divided into two: "supervised learning" and "unsupervised learning" to the pattern recognition and machine learning community. Classification, decision trees, neural networks, and regression analysis are well known supervised learning methods. In recent years regression analysis has drawn much attention for its wide range of applicability to the user of learning community. Regression can be performed using many different types of techniques, including classification, genetic algorithms, neural networks and optical flow estimation.

Supervised learning requires input data that has both predictor (independent) variables and target (dependent/response) variable, whose value is to be estimated. By various means, the process "learns" how to model (predict) the value of the target variable based on the predictor variables. Regression analysis is one of the most important supervised learning methods that have the objective to specify a relationship between the target and explanatory (independent) variables. Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given. Data contamination is a practical problem, introduced either due to unreliable data sources or during data transmission. Noise (data contamination) can greatly impair data learning. It is often more severe problem in the context of streams (large and high dimensions) because it interfaces with change adaptation. Regression analysis is not an exception to this kind of problem and regression diagnostics is a remedy. Problem is that most of the identification techniques are performed well only for small and few dimensional data.

In this paper we have a brief introduction of outliers in large and high-dimensional data in Section II, Section III focused on regression diagnostics. We propose a new diagnostic measure for identifying multiple influential observations for high dimensional large data set in Section IV. Numerical demonstrations are thereafter followed by conclusion in Section VI.

II. Outliers Detection in Large and High Dimensional Data

Outlier detection has been suggested for numerous applications, such as credit and fraud detection, clinical trials, voting irregularity analysis, network intrusion, severe weather prediction, geographic information system, and other data mining tasks [1], [2]. Outlying observations do not inevitably 'perplex' or 'mislead'; they are not necessarily 'bad' or 'erroneous', and the experimenter may be tempted in some situations not to reject an outlier but to welcome it as an indication of new and important findings.

Outlier detection methods can be divided into two, univariate and multivariate methods. Another fundamental taxonomy of outlier detection methods is between parametric and nonparametric methods that are model-free (*e.g.*, [3]). Statistical parametric methods either assume a known underlying distribution of the

observations or, at least, they are based on statistical estimates of unknown distribution parameters. These methods flag as outliers those observations that deviate from the model assumptions. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution [4]. Within the class of non-parametric methods one can set apart the data mining methods, also called distance-based methods. These methods are usually based on local distance measures and are capable of handling large databases (*e.g.,* [3], [5], [6]). Another class of outlier detection methods is founded on clustering techniques, where a cluster of small sizes can be considered as clustered outliers [7], [8].

III. Regression Diagnostics

The purpose of regression (model function) is to map a data item to a real valued function prediction variable, classical model is

$$Y = X\beta + \varepsilon . \quad (1)$$

We predict the output Y with least squares method via the model,

$$\hat{Y} = X\hat{\beta} \text{ (in vector-matrix form);} \quad (2)$$

where $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$ and the corresponding residual vector, $r = Y - \hat{Y} = (I - H)\varepsilon$; in scalar form,

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j ; \quad i = 1,2,...,n$$

where $H = X(X^T X)^{-1} X^T$ is the leverage matrix and its jth diagonal element is h_{ij}. Usually r_i and h_{ij} along with their different transformed versions are used for the diagnosis of outliers and leverage points respectively.

In regression, outliers can deviate into three ways (i) the change in the direction of response (Y) variable (ii) the deviation in the space of explanatory variable(s), deviated points in X-direction called leverage points, and (iii) the other is change in both the directions. It is known that as with outliers, high leverage points need not be influential, and influential observations are not necessarily high-leverage points. According to Belsley *et al.* [9], influential observation is one which either individual or together with several other observations has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations. Robust regression and regression diagnostics are two complementary approaches to study with unusual observations. Robust regression first wants to fit a regression to the majority of the data and then to discover outliers. On the contrary, field diagnostics is a combination of graphical and numerical tools. It is designed to detect and delete (if necessary recheck/correct) and then fit the good data by classical methods. It tries to iteratively detect possibly wrong data and reject/refit them through analysis of globally fitted model.

A large body of literature is now available (see [9-12]) for the identification of outliers and influential observations. Among them Cook's distance [13], DFFITS and DFBETAS [9] become very popular with the practitioners for identifying influential observations. These are based on single case deletion and unable to cope with multiple unusual observations. There are some group deletion measures (see [14]) but they are failed in case of high dimensional large data sets.

High dimensional data introduces several problems to traditional statistical analysis. Computation time increases more rapidly with number of explanatory variable than the number of observations. Another problem for group deletion diagnostics is to find group of outlying observation *d*, the deletion of which produces the largest reduction to the residual sum of squares.

IV. New Diagnostic Measure

We introduce a two-step diagnostic measure originated from Pena's [15] idea attaching with formal and/or informal group deletion methods (see [14]). Group deletion helps us to reduce the maximum disturbance by deleting the suspect group of influential cases at a time and 'the deletion of which produces the largest reduction in the residual sum of squares. It helps to make the data more homogeneous. We look how every point in a sample is influenced by the specific (*i*th) observation. In our method, after deleting the suspect group at a time we again delete specific sample point from each of the observations one after another and measure the difference between the forecast with and without the specific one. Group deletion attaching with Pena's [15] idea improves the measure and show nice results in presence of masking and/or swamping phenomena. Graphical displays like index plot and character plot of explanatory and response variables could give us some ideas about the influential observations, but these plots are not useful for higher dimension of regressors. There are some suggestions in the literature for using robust regression techniques like LMS, LTS, reweighted least squares (RLS) (see [10]) for finding the group of suspect influential observations. We consider one of the two proposed procedures of Hadi and Simonoff [14], where they suggested dividing the data set in two initial subsets: a 'basic' subset that contains the first k+1 clean observations and a 'nonbasic' subset that contains the remaining observations.

At the first step we find out all suspect influential (unusual) cases. We suggest using graphical display and/or regression diagnostic measures and/or robust regression methods

whatever can helps to find the suspect group with maximum number of influential cases.

In the second step, we assume that *d* observations among a set of *n* observations are suspected as influential observations and to be deleted. Let us denote a set of cases 'remaining' in the analysis by *R* and a set of cases 'deleted' by *D*. Hence without loss of generality, assume that these observations are the last *d* rows of *X* and *Y* so that

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \qquad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}.$$

After formation of the deletion set indexed by *D*, we would compute the fitted values $\hat{Y}^{(-D)}$. Let $\hat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by *D* is omitted. We define the vector of difference between $\hat{y}_j^{(-D)}$ and $\hat{y}_{j(i)}^{(-D)}$ as

$$t_{(i)}^{(-D)} = (\hat{y}_1^{(-D)} - \hat{y}_{1(i)}^{(-D)}, ..., \hat{y}_n^{(-D)} - \hat{y}_{n(i)}^{(-D)})^T \quad (3)$$

$$= (t_{1(i)}^{(-D)}, ..., t_{n(i)}^{(-D)})^T \quad (4)$$

and

$$t_{j(i)}^{(-D)} = \hat{y}_j^{(-D)} - \hat{y}_{j(i)}^{(-D)} = \frac{h_{ji}\hat{\varepsilon}_i^{(-D)}}{1-h_{ii}}, j = 1,2,...,n \quad (5)$$

where

$h_{ji} = x_j^T (X^T X)^{-1} x_i$ and $\hat{\varepsilon}_i^{(-D)} = y_i - \hat{y}_i^{(-D)}$.

Finally, we introduce our new measure as squared standardized norm,

$$M_i = \frac{t_{(i)}^{(-D)^T} t_{(i)}^{(-D)}}{kV(\hat{y}_i^{(-D)})}, \quad (6)$$

where

$$V(\hat{y}_i^{(-D)}) = s^2 h_{ii} \qquad and \qquad s^2 = \frac{\hat{\varepsilon}^{(-D)^T}\hat{\varepsilon}^{(-D)}}{n-k}.$$

Using (3) to (6), we obtain

$$M_i = \frac{1}{ks^2 h_{ii}} \sum_{j=1}^{n} h_{ji}^2 \frac{\hat{\varepsilon}_i^{(-D)2}}{(1-h_{ii})^2}. \quad (7)$$

We use a confidence bound type cut-off value for it, and consider the *i*th observation to be influential if it satisfies the rule

$$|M_i| > median(M_i) + 4.5MAD(M_i), \quad (8)$$

where

$$MAD(M_i) = median\{|M_i - median(M_i)|\}/0.6745.$$

V. Numerical Examples

We consider here two artificial large data sets, one is multi-structural clustered (step) and the other is high-dimensional, for the identification of influential observations. Such data sets are frequently used in classification, cluster analysis and optical flow estimation. We want to show the performance of our new measure and compare with the existing popular diagnostic methods.

A. Multi Structural Step Data

Our first example uses an artificial data set that is taken from (see [16]) a paper of optical flow estimation (computer vision). In this experiment, we investigate how a multistructural clustered data is analyzed which is contaminated by number of outliers (noise). We generate the step data set, since every line is corrupted by Gaussian noise with zero mean and different variances. The *i*th line has n_i samples, n_0 data points are randomly distributed in the range of (0,100). The data points are defined as follows.

Steps are: X= (0, 55), Y=30, n_1=120, σ =1.5 ('+');
X= (20,100), Y=60, n_2=50, σ =2 ('-'); n_0= 50 ('o').

First we plot the points with LS and robust (LMS and LTS) line, it is clearly shown that data sets are clustered on two lines and some observations do not hold any pattern. As the common definition, observations those are assigned 'o' and '-' in Fig.1 can be treated as outliers (apart from the bulk of the data). Judging from the LMS line it is worth mentioning that last 100s are outliers. Now we apply our newly proposed diagnostic technique and form the deletion set *D* by the hundreds. Index plots (Fig. 2. (b) and (d)) of residuals and leverage points do not show any indication of heterogeneity.

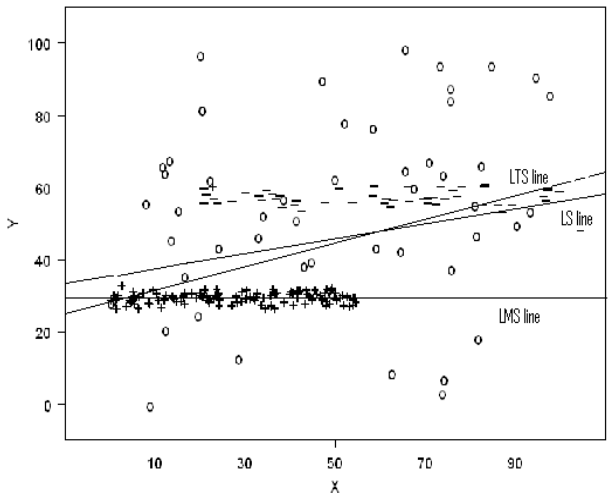

Fig.1. Scatter plot of multi structural step data.

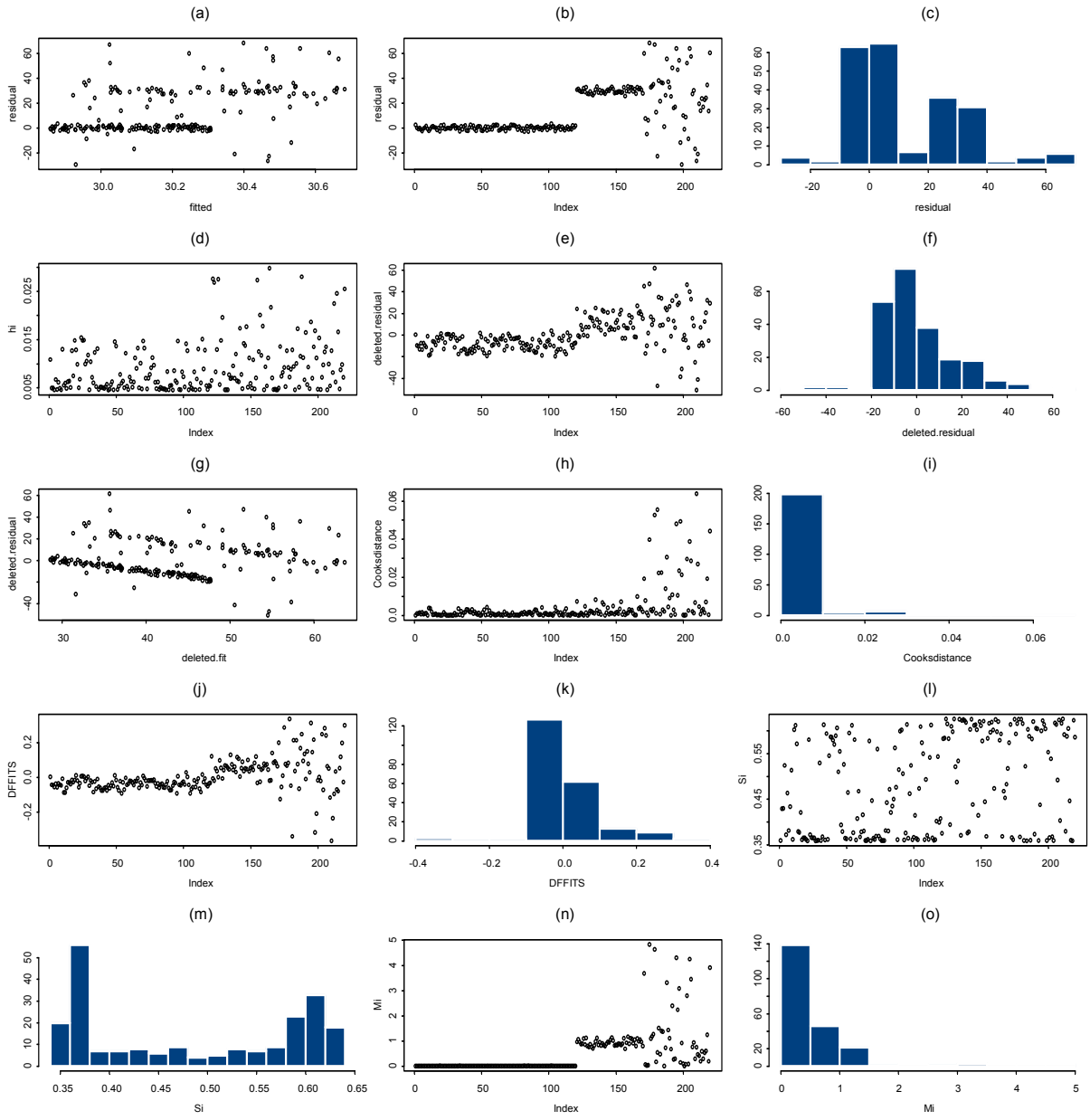

Fig. 2. Regression diagnostics for multistructural step data; (a) Plot of residual versus fitted value (b) Index plot of residuals (c) Histogram of residuals (d) Index plot of leverage, hi values (e) Index plot of deleted residuals (f) Histogram of deleted residuals (g) Deleted residual versus deleted fit (h) Index plot of Cook's distance (i) Histogram of Cook's distance (j) Index plot of DFFITS (k) Histogram of DFFITS (l) Index plot of Si (Pena) (m) Histogram of Si (n) Index plot of Mi and (o) Histogram of Mi.

We compute the diagnostic measures Cook's distance and DFFITS along with Pena's measure (S_i) and M_i (new measure) and use them to draw index (dot) plot and histogram. Without Mi, rest 3 are failed to indicate influential cases properly. Index plot of Mi, Fig. 2 (n) shows that last 100 observations are influential with a few masking observations. Histogram of Mi, Fig. 2 (o) clearly indicates the presence of three different patterns of observations.

B. Large and High Dimensional Data

We consider an artificial data set that is generated in a similar fashion described by Pena [15]. The data set shows the presence

of heterogeneous variances in the sample points and a mixture of continuous and categorical variables, generated by the model

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{20} X_{20} + \beta_{21} Z + \varepsilon . \quad (9)$$

We generate 500 observations, where X's have 20 dimensions and they are independent random drawings from uniform distributions. The first 400 observations for each of the X variables are generated from Uniform (0, 10) and last 100 observations (401-500) from Uniform (9, 10) that makes the presence of heterogeneous variance in the data set. For the categorical explanatory variable Z, the first 400 observations are set at $z = 0$ and the last 100 observations are set at $z = 1$. For the null model we generate errors from Normal (0, 1). The parameter values have been chosen as $\beta_0 = \beta_1 = ... = \beta_{20} = 1$, and β_{21} = – 100 so that the standard diagnostics do not show any evidence of heterogeneity.

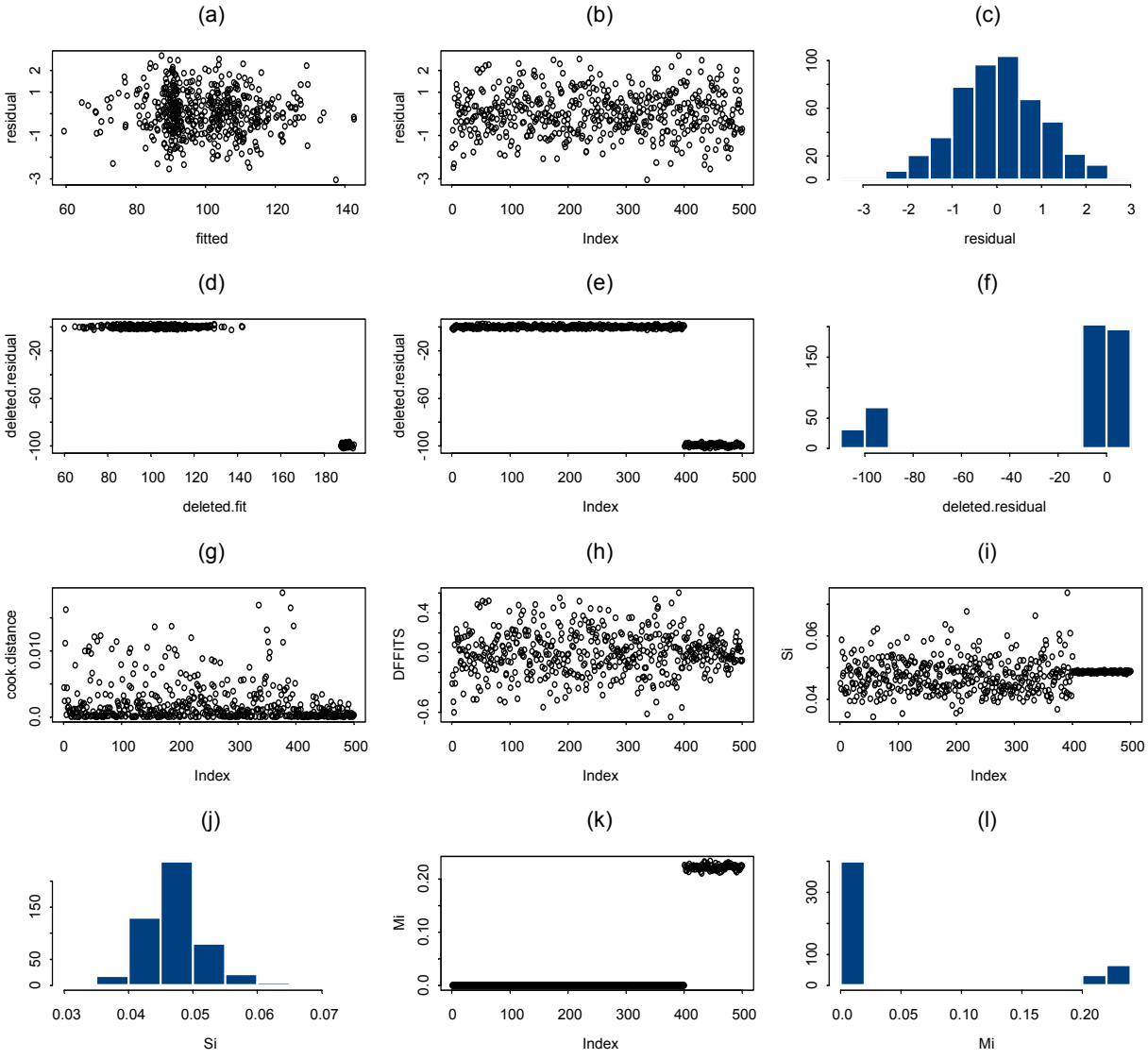

Fig. 3. Influence analysis of the large and high dimensional data; (a) Plot of residual versus fitted value (b) Index plot of residuals (c) Histogram of the residuals (d) Plot of deleted residual versus deleted fitted value (e) Index plot of deleted residuals (f) Histogram of the deleted residuals (g) Index plot of Cook's distance (h) Index plot of DFFITS (i) Index plot of Si (j) Histogram of Si (k) Index plot of Mi, and (l) Histogram of Mi.

Results that we obtain from this artificial data are shown in Fig. 3. The residual versus fitted value plot, Fig. 3 (a)

shows no indication of heterogeneity. When the last 100 observations are deleted, the deleted residual versus deleted fitted value plot and histogram of deleted residual (Fig. 3. (d), (f)) show clear indications of the presence of heterogeneity. We consider above mentioned three measures to identify the 100 influential cases but their index plots (Fig. 3. (g), (h), and (i)) show they are totally failed to identify the influential cases. In index plot, Fig. 3. (k) we see that proposed M_i can successfully identify all influential cases. Fig. 3. (l), histogram of M_i also shows the heterogeneous variances in the data set that is absent in histogram of S_i.

VI. CONCLUSION

In this article, we propose a group deletion diagnostic measure for identifying influential observations (outliers) in linear regression that performs well for large and high dimensional data. The technique also gives some clear indications of presence of influential cases in a multidimensional step (cluster) data when the existing popular diagnostic methods are not well enough. To get real data involving thousands of cases and hundreds of dimensions is not easy and also expensive. Time and space constrains do not permit us to perform the tasks this time; we shall try to do it as a further work.

ACKNOWLEDGEMENT

The authors are thankful to the anonymous reviewers for their valuable suggestions and to the organizing and publication committee of the conference.

REFERENCES

[1] T. Fawcett, and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, Vol.1, No. 3, pp. 291-316, 1997.

[2] K. L. Penny, and I. T. Jolliffe, "A comparison of multivariate outlier detection methods for clinical laboratory safety data," *The Statistician*, Vol. 50, No. 3, pp. 295-308, 2001.

[3] G. L. Williams, R. A. Baxter, H. X. He, S. Hawkins, and L. Gu, "A comparative study of RNN for outlier detection in data mining," *IEEE International Conference on Data Mining (ICDM'02)*, Maebashi, Japan, 2002.

[4] S. Papadimitriou, H. Kitawaga, P. G. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," *International. Research Laboratory Technical report*, No. IPR=TR-02-09, 2002.

[5] S. Hawkins, H. X. He, G. J. Williams, and R. A. Baxter, "Outlier detection using replicator neural networks," in Proceedings of the *Fifth International Conference on Data Warehousing and Knowledge Discovery*, France, 2002.

[6] E. Knorr, R. Ng, and V. Tucakov, "Distance based outliers: algorithms and applications," *VLDB Journal: Very Large Data Bases*, Vol. 8(3-4), pp. 237-253, 2000.

[7] E. Acuna, and C. Rodriguez, "A meta analysis study of outlier detection methods in classification," Technical Paper, Department of Mathematics, University of Puerto Ricoat Mayaguez, http://www.academic.uprm.edu/~eacuna/paperout.pdf, in Proceedings, IPSI, 2004.

[8] S. Shekhar, C. T. Lu, and P. Zhang, "Detecting graph based spatial outlier," *Intelligent Data Analysis: An International Journal*, Vol. 6, No. 5, pp. 451-468, 2002.

[9] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: J. Wiley & Sons, 1980.

[10] P. J. Rousseeuw, and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: J. Wiley & Sons, 1987.

[11] S. Chatterjee, and A. S. Hadi, *Sensitivity Analysis in Linear Regression*, New York: J. Wiley & Sons, 1988.

[12] A. C. Atkinson, and M. Riani, *Robust Diagnostic Regression Analysis*, New York: Springer, 2000.

[13] R. D. Cook, "Detection of influential observations in linear regression", *Technometrics*, Vol.19, pp.15-18, 1977.

[14] A. S. Hadi, and J. S. Simonoff, "Procedure for the identification of outliers in linear models," *Journal of American Statistical Association*, Vol. 88, pp. 1264 - 1272, 1993.

[15] D. Pena, "A new statistic for influence in linear regression," *Technometrics*, Vol. 47, pp.1-12, 2005.

[16] J. Colliez, F. Dufrenois, and D. Hamad, Robust regression and outlier detection with SVR: application to optical flow estimation, http:// www. macs.hw.ac.uk/bmvc2006/papers/446.pdf, access 8th March, 2008.