Regression Diagnostics for Multiple Model Step Data

Preprint · October 2009

CITATIONS
0

READS
10

2 authors:

Abdul Awal Md Nurunnabi
University of Luxembourg
94 PUBLICATIONS **1,483** CITATIONS
SEE PROFILE

M. Nasser
University of Rajshahi
33 PUBLICATIONS **303** CITATIONS
SEE PROFILE

Regression Diagnostics for Multiple Model Step Data

A. A. M. Nurunnabi
School of Business, Uttara University
Dhaka-1230, Bangladesh
e-mail: pyal1471@yahoo.com

Mohammed Nasser
Department of statistics, University of Rajshahi
Rajshahi-6205, Bangladesh
e-mail: mnasser.ru@gmail.com

***Abstract*—In many vision and image problems there are multiple structures in a single data set and we need to identify the multiple models. To preserve most structures in presence of noise makes the estimation difficult. In such case for each structure, data which belong to other structures are also outliers in addition to the outliers for all the structures. Robust regression techniques are commonly used to serve the model building process for noisy data to the vision community, that fits the majority data and then to discover outliers, they tend to fail to cope with the situation. In this paper we show, a newly proposed regression diagnostic measure is capable for identifying large fraction of outliers, and regression diagnostics may be a better choice to the robust regression. We demonstrate the whole thing through several artificial multiple model step data.**

***Keywords- cluster analysis; computer vision; image analysis; multiple structural data; outlier; regression diagnostics; robust regression*.**

I. INTRODUCTION

Fitting a regression model to observation data is a basic statistical technique. Much of computer vision and image analysis tasks involve the extraction of meaningful information from images using concepts akin to regression. Tasks include: automated surveillance, film restoration, motion estimation, object recognition, optical flow estimation, robot vision, video coding, visualization, etc. Many of them contain data set with multiple structures, and due to noise and discrete nature of digitalized images all image analysis methods so far are subject to errors. Robust regression and regression diagnostics are two complementary approaches to study with outlier contaminated data. Robust regression methods are commonly used to serve the purpose to the vision community (*e.g.*, [5], [6], [14], [19], [21]). The problems of estimation in robust regression receive a lot of attention in computer vision literature [22], [23]. These techniques focus on the estimation of a single model and typically differ in their assumptions, efficiency and capability of handling different portion of outliers. Robust regression aims to fit a model from the majority of the data observations. They can tolerate gross outliers [12], [18]. Wang [21] mentioned that, they can breakdown at unexpectedly lower percentage when the outliers are clustered; also, they can not tolerate more than 50% outliers. Hillebrand and Muller [11] pointed out that, the ability to remove large amount of noise and the ability to preserve most structures are desirable properties to study in image analysis. These goals are difficult to achieve at the same time and, in particular, it is difficult to remove (refit if necessary) outliers and to preserve discontinuities in multiple model. This study presents: a new regression diagnostic measure can efficiently identifies outliers in a multiple structural data set, and regression diagnostics may be an alternative to robust regression.

This paper is arranged as follows: In Section 2, we give some general ideas of multistructural data and cluster analysis. In Section 3 we discuss, about regression diagnostics, a new diagnostic measure and the proposed algorithm. Section 4 provides empirical evaluation through two examples; we conclude in Section 5.

II. MULTI STRUCTURAL DATA AND CLUSTER ANALYSIS

This section gives us the basic ideas and problems in multistructural data to perform regression analysis and a very short discussion on cluster analysis.

A. Multiple structural Data

A data with multiple structures is characterized by the presence of different model (instance) with different set of parameters in the same. In a given image there are usually several populations of data. Some parts may correspond to one object in a scene and other parts will correspond to other, rather unrelated, objects. In presence of multiple structures in a same data set, for each structure, data which belong to other structures are also outliers (pseudo outliers) in addition to the true outliers (gross outliers). Moreover, the outliers due to faulty feature extraction, sensor noise, segmentation errors, etc. attach with the pseudo outliers. According to Suter and Wang [19], it will rarely happen that a given population achieves the critical size of 50% of the total population and, therefore, techniques that have been touted for their high breakdown point [9] are no longer reliable candidate from this point of view. "The presence of structural (pseudo) outliers is the most damaging for the performance of robust estimators and almost always the breakdown condition is due to such data [20]".

Regression analysis studied usually has a single dominant population with some percentage of outliers (gross outliers). So it does not consider pseudo outliers in its

estimation process. As the consequence, we need to partition the multiple structural data before performing regression analysis. Cluster analysis can helps us to remove the problem of pseudo outliers by clustering the data into homogeneous groups.

B. Cluster Analysis

Cluster analysis or data segmentation has a variety of goals. All relate to grouping of objects (observations) into subsets or "clusters" such that those within cluster are more closely related to one another than objects assigned to different clusters [10]. Clusters are natural groupings of data items based on similarity matrices or probability density models. Clustering pertains to unsupervised learning, when data with class levels are not available. Mitra and Acharya [15] mentioned that, clustering is very important in multimedia data mining, in order to cluster similar multimedia contents together for efficient indexing and storage in multimedia databases. In two or three dimensions, clusters can be visualized and more than three dimensions we need some kind of analytical assistance. Clustering algorithm fall into two categories: partitioning algorithm and hierarchical algorithm. One of the most popular partitioning algorithms is 'partitioning around medoids' (PAM), uses the most centrally located object in a cluster, the medoid, instead of the mean. It is closest to the corresponding mean. The basic steps are outlined as follows: i) arbitrarily choose *c* objects as the initial medoids, ii) assign each remaining data object (pattern) to the cluster for the closest medoid, iii) replace each of the medoids by one of all the nonmedoids (causing the greatest reduction in square error), as long as the quality of clustering improves, and iv) iterate until the criterion function converges. To serve our purpose we use PAM because, it is more robust (It minimizes a sum of dissimilarities instead of a sum of squared Euclidian distance.), and it provides graphical displays.

III. Regression Diagnostics and The New Measure

The usual form of a linear regression model is

$$Y = X\beta + \varepsilon, \qquad (1)$$

where Y is an $n \times 1$ vector of response, X is an $n \times k$ ($n > k; k = p + 1$) full rank matrix of explanatory variables including one constant explanatory variable, β is a $k \times 1$ vector of unknown finite parameters and ε is an $n \times 1$ vector of i.i.d. random disturbances each of which follows $N(0, \sigma^2)$. The most common method of predicting the output Y is the well known least squares (LS) method via the model,

$$\hat{Y} = X\hat{\beta} \text{ (vector-matrix form)}, \qquad (2)$$

where $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$ and the corresponding residual vector, $r = Y - \hat{Y} = (I - H)\varepsilon$; where $H = X(X^T X)^{-1} X^T$ is the leverage matrix. In scalar form, $r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$, $i = 1,2,...,n$ and h_{ii} is the ith diagonal element of H. LS is the most common method mainly because of computational simplicity and achieves some optimum properties with some underlying assumptions [3]. However, this method is extremely sensitive to outliers and its breakdown point is 0, *i. e.*, only one single outlier can totally destroy its results.

A. Regression Diagnostics

Regression diagnostics is a combination of graphical and numerical tools. It is designed to detect and delete (or refit) the outliers first and then to fit the good data by classical (least squares) methods. According to Huber [13], robustness and diagnostics are complementary approaches to the analysis of data any one of the two is not good enough. Hence the basic need of regression diagnostics is the identification of outliers. A large number of diagnostic methods are existed in the literature (*e. g.*, [1], [2], [3], [7]).

B. Newly Proposed Diagnostic Measure

We introduced a two-step group deletion diagnostic measure (in [16], [17]). 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. This measure based on group deletion, shows nice results in presence of masking and/or swamping [8] 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 [18]) for finding the group of suspect influential observations. We consider one of the two proposed procedures of Hadi and Simonoff [8], 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 $V(\hat{y}_i^{(-D)}) = s^2 h_{ii}$, and $s^2 = \frac{\hat{\varepsilon}^{(-D)^T}\hat{\varepsilon}^{(-D)}}{n-k}$.

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

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

C. The Proposed Algorithm

The complete algorithm can be described as follows:

Step 0: Input a data set.
Step 1: Graph the data and try to find the number of cluster.
Step 2: Use PAM to get the clusters.
Step 3: Apply LMS or appropriate robust or diagnostic method to all the clusters and find the suspect outliers.
Step 4: Make deletion set *D* by the suspect outliers and apply the new diagnostic measure.
Step 5: Identify the final outliers by the proposed measure, and fit the models by LS.

IV. Experiments and Analysis

In order to assess the proposed algorithm, we carry out experiments through two artificial multi level step data sets that have been used in vision community.

A. Multiple Line Data

Our first example uses a data set like Colliez [6], used in computer vision literature. We generate the data set seems two lines, since each line is corrupted by Gaussian noise with zero mean and different variance. The *i*th line (structure) has n_i samples, n_0 data points are randomly distributed in the range of (0-100). Multiple structures are constructed as follows:

Line 1: $x : (0-55), y = 30, n_1 = 120, \sigma = 1.5$;

Line 2: $x : (20-100), y = 60, n_1 = 100, \sigma = 2.0$;

Gross outliers: $x : (0-100), y = (0-100), n_0 = 30$.

The outliers (gross + pseudo) percentage for *i*th line is defined as ε_i, $\varepsilon_1 = 52(12+40)\%$ and $\varepsilon_2 = 60(12+48)\%$.

From Fig. 1 (a), we see, because LMS has only a 0.5 breakdown point, it can not tolerate more than 50% outliers, LMS line performs only for first line and ignores rest of the observation because first line contains 120 observations which is more than half (100; 2nd line), without the gross outliers. LS line considers all 250 observations and makes a fun. We use PAM as a cluster algorithm; we get two clusters (see Fig. 1 (b)); one for (1-120) and another for (121-250). If we look at Fig. 1 (c), LMS performs well but till now LS is confusing, it fails to fit second line. Applying LMS, we get 3 (39, 52, 57) and 28 (101-117, 119-121, 123-130) observations are outliers for line 1 and 2 respectively. Now we apply our newly proposed diagnostic technique and form the deletion sets *D* by the 3 and 28 observations for line 1 and line 2 respectively. We see from Fig. 1 (index plots of Mi; *d*, and *e*) that some observations are above the cutoff line and identified as influential, plots also show that there are some masking effects. Seeing at Fig. 1 (f) we can say, the newly proposed algorithm properly identifies two models in a single data set.

B. Three Stairs Data

This example generates a data set as Chen *et al.*, [5], it looks 3 stairs (steps) one after another. Each stair is corrupted by Gaussian noise with mean zero and unit variance. All the stairs have equal number of observations (n_i). Ten single outlier data points are randomly distributed between 0 and 100. Stairs are as follows:

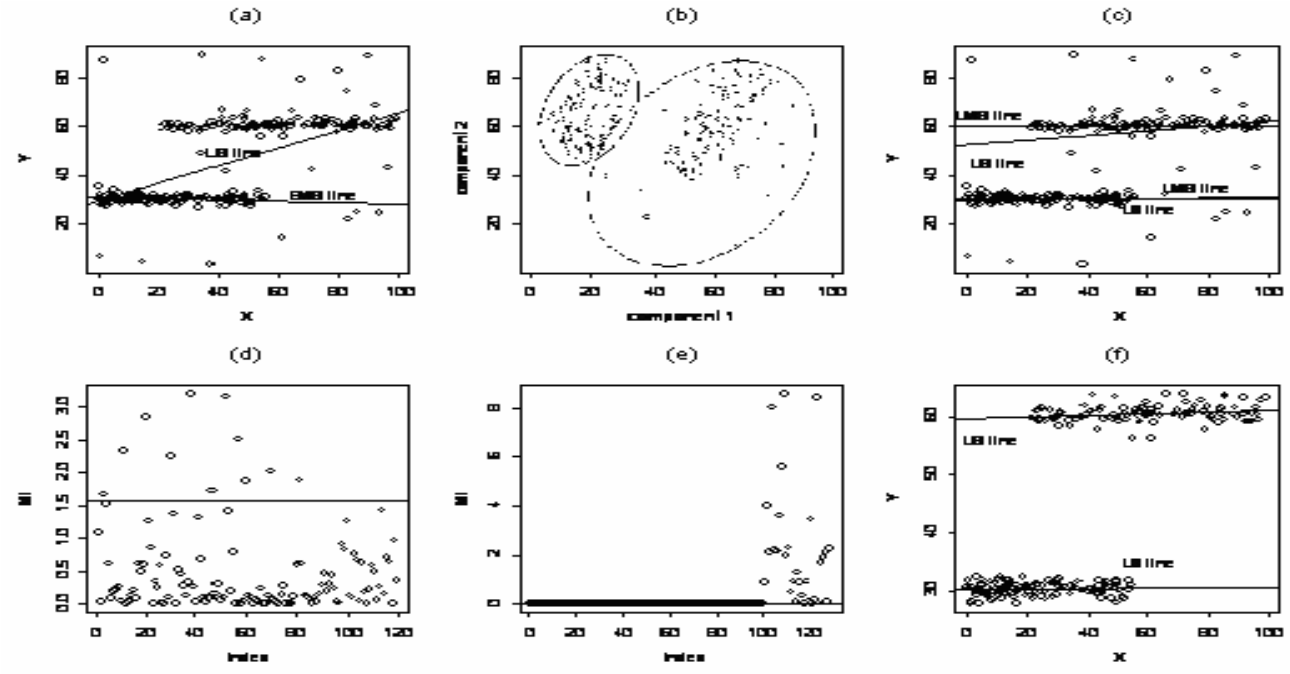

Figure 1. Multiple line data; (a) Scatter plot with LS and LMS line, (b) Clustering by PAM, (c) Scatter plot with fitted lines after clustering, (d) Index plot of Mi for 1st structure, (e) Index plot of Mi for 2nd structure, and (f) Scatter plot with LS fitted lines without outliers.

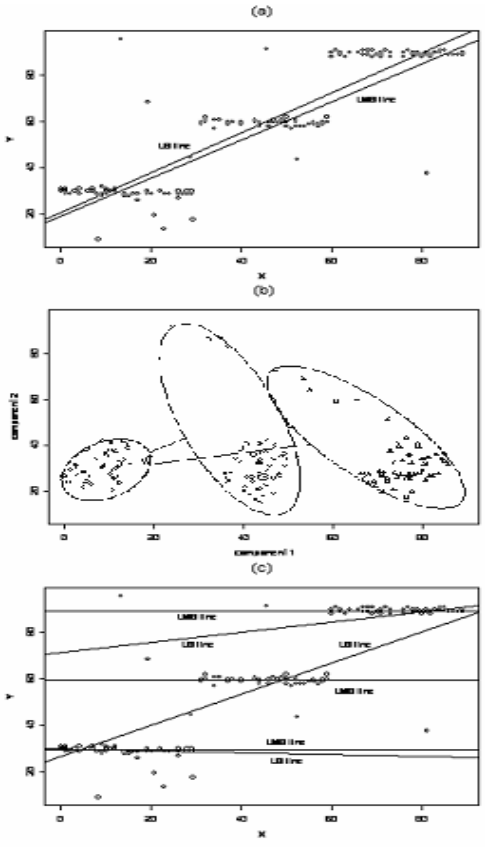

Figure 2. Three stairs data; (a) scatter plot with LS and LMS lines (b) Three clusters' plot (c) Scatter plot with LS and LMS line.

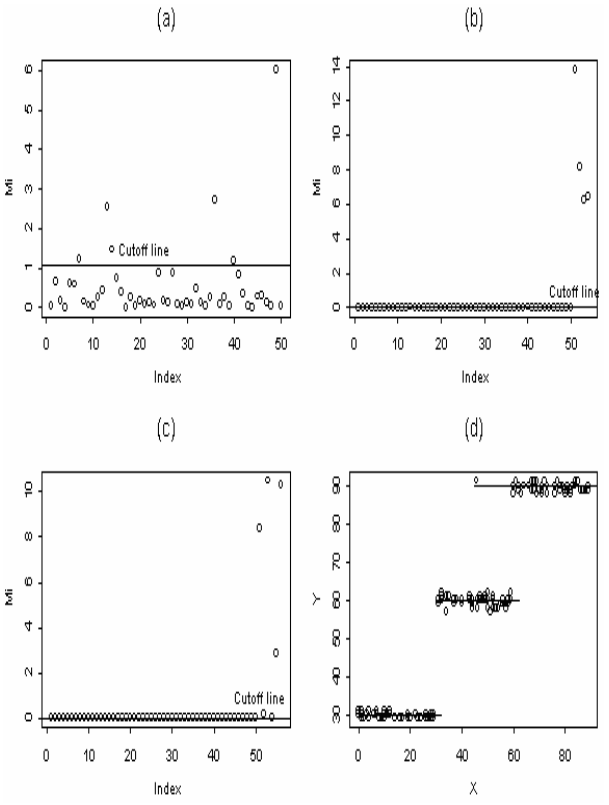

Figure 3. Three stairs data; (a) Index plot of Mi for 1st cluster, (b) Index plot of Mi for 2nd cluster, (c) Index plot of 3rd cluster (d) Scatter plot and 3 stairs by LS after outlier deletion.

Stair 1: $x:(0-30), y=30, n_1=50$.

Stair 2: $x:(30-60), y=60, n_2=50$

Stair 3: $x:(60-90), y=90, n_3=50$

Gross outliers: $x:(0-100), y=(0-100), n=10$.

Outliers' contamination rates are same for all, 68.75%.

As usual we perform the scatter plot to assess the underlying behavior in the data set. Fig. 2 (a) shows the LS and LMS lines are almost same in directions. Both are erroneous, simply by eye judgment. The scatter plot indicates 3 steps to judge. We use PAM as predefined cluster algorithm and find 3 clusters properly. We know in prior, minimum 10 outliers are present in the data set. LS must falsify the estimation. We use robust technique (LMS) to identify the gross outliers. Applying LMS, we get 4 (152, 154, 156, and 159) and 6 (151, 153, 155, 157, 158, and 160) cases as outliers in 2nd and 3rd clusters. Fig. 2 (c) shows that although LMS performs well in 3 steps but LS fails to fit for the 2nd and 3rd lines. Now we form the deletion set *D* by the 4 and 6 observations for the last two steps and apply the new measure. Fig. 3 (a, b, and c) show that our newly proposed diagnostic measure (Mi) successfully identifies all the outliers even in first cluster, which was masked before, LMS did not find them (cases 7, 13, 14, 36, 40, and 49). Index plots of the Mi, Fig. 3 (b, c) properly identify the gross outliers and coincide with the LMS results. Moreover, Fig. 3 (d) justifies the proper stairs fitting and identification tasks as well.

V. Conclusion

In this paper we develop a new algorithm to perform regression diagnostics in multiple model step data. Paper shows the newly proposed diagnostic measure correctly identifies outliers for multiple structural data. Experiments in the paper show that the proposed method successfully extracts different structures (model/instances) in spite of high noise level (more than 50%). By segmenting the data set into some clusters the proposed method can keep it away from the problems of large data analysis. The proposed technique could be used as a better alternative to robust regression in image analysis and computer vision.

References

[1] A. C. Atkinson, and M. Riani, Robust Dignostic Regression Analysis, New York: Springer, 2000.

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

[3] S. Chatterjee, and A. S. Hadi, Sensitivity Analysis in Linear regression, New York: Wiley & Sons, 1988.

[4] S. Chatterjee, and A. S. Hadi, Regression Analysis by Examples, New York: Wiley & Sons, 2006.

[5] H. Chen, P. Meer, and D. Tyler, "Robust Regression for Data with Multiple Structures," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, (Kauai, Hi), Dec. 2001, pp. 1069-1107.

[6] J. Colliez, F. Dufrenois, and D. Hamad, Robust Regression and Outlier Detection with SVR: Application to Optical Flow Estimation, 2006, http://www.macs.hw.ac.uk/bmvc2006/papers/446.pdf.

[7] R. D. Cook, and S. Weisberg, Residuals and Influence in Regression, New York: Chapman and Hall, 1982.

[8] A. S. Hadi, and J. S. Simonoff, "Procedures for the Identification of outliers in Linear Models," Journal of American Statistical Association, Vol. 88, 1993, pp.1264-1272.

[9] F. R. Hampel, "A General Qualitative Definition of Robustness," Annals of Mathematical Statistics, Vol. 42, 1971, pp. 1887-1896.

[10] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning; Data Mining, Inference and Prediction, New York: Springer, 2001.

[11] M. Hillebrand, and C. H. Muller, "Outlier Robust Corner-Preserving Methods for Reconstructing Noisy Images," The Annals of Statistics, Vol. 35, 2007, pp.132-165.

[12] P. J. Huber, Robust Statistics, New York: Wiley & Sons, 1981.

[13] P. J. Huber, Between robustness and diagnostics, In W. Stahel, and S. Weisberg, editors, Direction in Robust Statistics and Diagnostics, Part-1, New York: Springer, 1991, pp. 121-130.

[14] P. Meer, D. Mintz, and A. Rosenfeld, "Robust Regression Methods for Computer Vision: A Review," International Journal of Computer Vision, Vol. 6 (1), 1991, pp. 59-70.

[15] S. Mitra, and T. Acharya, Data Mining: Multimedia, Soft Computing, and Bioinformatics, Wiley & Sons (Asia) Pte. Ltd. 2004.

[16] A. A. M. Nurunnabi, Robust Diagnostic Deletion Techniques in Linear and Logistic Regression, M. Phil. Thesis, Unpublished, Rajshahi University, Bangladesh, 2008.

[17] A. A. M. Nurunnabi, and M. Nasser, "Regression Diagnostics in Large and High Dimensional Data," Proc. of International Conference of Computer and Information Technology, Bangladesh, Dec. 2008, in Press.

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

[19] D. Suter, and H. Wang, Robust fitting using mean shift: applications in computer vision, In M. Hubert, G. Pison, A. Struyf, and van Aelst, editors, Theory and Applications of Recent Robust Methods, Statistics for Industry and Technology, Birkhauser, Basel, 2004, pp. 307-318.

[20] D. E. Tyler, "Breakdown Properties of M-estimators of Multivariate Scatter," Technical report, Statistics Department, Rutgers University, New Jersey, Piscataway , July 1986.

[21] H. Wang, Robust Statistics for Computer Vision: Model Fitting, Image Segmentation, and Visual Motion Analysis, Ph. D. Thesis, Unpublished, Monash University, Australia, 2004.

[22] H. Wang, and D. Suter, "A Novel Robust Method for Large Number of Gross Errors," Proc. of 7th International Conference on Control, Automation, Robotics and Vision, Singapore, 2002, pp. 326-331.

[23] H. Wang, and D. Suter, "LTSD: A Highly Efficient Symmetry Based Robust Estimator," Proc. of 7th International Conference on Control, Automation, Robotics and Vision, Singapore, 2002, pp. 332-337.