Pakistan Journal of Statistics and Operation Research

Identification of High Leverage Points in Linear Functional Relationship Model

Abu Sayed Md. Al Mamun[1*], A. H.M. Rahmatullah Imon[2] , Abdul Ghapor Hussin[3], Yong Zulina Zubairi[4], Sohel Rana[5]

* Corresponding Author

1. Abu Sayed Md. Al Mamun, Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh., mithun_stat@yahoo.com
2. A. H.M. Rahmatullah Imon, Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA, imon_ru@yahoo.com
3. Abdul Ghapor Hussin, Faculty of Science and Defence Technology, National Defence University of Malaysia, Kuala Lumpur, Malaysia, ghapor@upnm.edu.my
4. Yong Zulina Zubairi, Mathematics Division, Centre for Foundation Studies in Science, University of Malaya, Kuala Lumpur, Malaysia, yzulina@um.edu.my
5. Sohel Rana, Department of Mathematical and Physical Sciences, East West University, Dhaka, Bangladesh, srana@ewubd.edu

Abstract

In a standard linear regression model the explanatory variables, X , are considered to be fixed and hence assumed to be free from errors. But in reality, they are variables and consequently can be subjected to errors. In the regression literature there is a clear distinction between outlier in the Y - space or errors and the outlier in the X-space. The later one is popularly known as high leverage points. If the explanatory variables are subjected to gross error or any unusual pattern we call these observations as outliers in the X - space or high leverage points. High leverage points often exert too much influence and consequently become responsible for misleading conclusion about the fitting of a regression model, causing multicollinearity problems, masking and/or swamping of outliers etc. Although a good number of works has been done on the identification of high leverage points in linear regression model, this is still a new and unsolved problem in linear functional relationship model. In this paper, we suggest a procedure for the identification of high leverage points based on deletion of a group of observations. The usefulness of the proposed method for the detection of multiple high leverage points is studied by a well-known data set and Monte Carlo simulations.

Key Words: Errors in variable, Leverages, Masking, Swamping, Monte Carlo simulation

Mathematical Subject Classification: 62J05 and 62J20.

1. Introduction

The linear functional relationship model (LFRM) is an extension of a linear regression model (LRM) which allows for sampling variability in the measurements of both the response and explanatory variables. In regression the model is poorly fitted because of the presence of outliers. It is a common practice over the years to use residuals for the identification of outliers. Residuals are in fact estimates of the true errors that occur in the *Y*-space. We anticipate at this point that fitting of the LFRM could be even more complicated because here outliers could occur in the *X*-space more frequently than the linear regression model. Outliers in the *X*-space are called high leverage points in the regression literature since they exert too much weight on the fitting of the model. When we use the ordinary least squares (OLS) or the maximum likelihood (ML) method for fitting a regression line, the resulting residuals are functions of leverages and true errors. Thus high leverage points together with large errors (outliers) may pull the fitted line in a way that the fitted residuals corresponding to those outliers might be too small and this may cause masking (false negative) of outliers. For the same reason the residuals corresponding to inliers may be too large and this may cause swamping (false positive). Peña and Yohai (1995) pointed out that high leverage cases are mainly

responsible for masking and swamping of outliers in linear regression. The unfortunate consequences of the presence of high leverage points in linear regression have been studied by many authors. The presence of a high leverage point could increase (often unduly) the value of R^2. Chatterjee and Hadi (1988) mentioned the existence of collinearity-influential observations whose presence could induce or break the collinearity structure among the explanatory variables. Kamruzzaman and Imon (2002) and Imon and Khan (2003a) pointed out that high leverage points may be the prime source of collinearity-influential observations. Imon (2009) pointed out that in the presence of high leverage points the errors not only become heteroscedastic, they might produce big outliers as well. Another way to deal with outliers is to use M-estimators (Mahdizadeh et al., 2020; Zamanzade et al.; 2020; Zamanzade et al., 2019 and Zamanzade et al., 2018). This could make the procedures for the detection of heteroscedasticity very complicated. That is why the identification of high leverage points is essential before making any kind of inference.

In this paper our main objective is to identify high leverage points in a linear functional relationship model. Although some efforts have been done on the identification of outliers and influential observations in LFRM e.g. (Abdullah,1995; Vidal, 2007; and Wellman, 1991), but so far as we know, there is no reported work in the identification of high leverage points in LFRM. Let us consider a simple linear regression model

$$y_i = \alpha + \beta X_i + \varepsilon_i \tag{1}$$

where y_i is the response, X_i is (supposed) explanatory variable assumed to be constant and specific assumption made on ε_i. We feel that the assumption of X_i being constant in model (1) may not appropriate in reality, instead we introduce a linear functional relationship model.

Consider the following model

$$y_i = Y_i + \varepsilon_i, \quad x_i = X_i + \delta_i,$$
$$\text{and} \qquad Y_i = \alpha + \beta X_i \quad \text{for } i = 1, 2, ..., n. \tag{2}$$

where the two linearly related unobservable variables *X* and *Y* are considered as the true part and the corresponding random variables *x* and *y* are observed with random errors δ_i and ε_i. The unobservable *X* and *Y* are fixed (nonstochastic) and (2) is called a functional relation. So the main difference between a LRM and a LFRM is that in LRM it is assumed that the explanatory variable is free from error but in LFRM it is subjected to error. In section 2, we discuss different measures of leverages. Since all conventional measures of leverages are based on fixed explanatory variable, in section 3 we introduce the estimated values of *X* in LFRM which can be considered as fixed. In section 4, we discuss a procedure of the identification of high leverage points in LFRM. The usefulness of this proposed measure is studied through real world data in section 5 and through Monte Carlo simulations in section 6.

2. Measures of Leverages

In regression analysis it is sometimes very important to know whether any set of *X*-values are exerting too much influence on the fitting of the model. According to Hocking and Pendleton (1983) "*high leverage points are those for which the input vector* $\mathbf{x_i}$ *, in some sense, far from the rest of the data."*

Let us consider a *k* variable regression model

$$\mathbf{Y} = \mathbf{X\beta} + \hat{\boldsymbol{\varepsilon}} \tag{3}$$

A set of influential *X*-values is known as a high leverage point. The OLS residual vector can be expressed in terms of the true disturbance vector as

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{W})\boldsymbol{\varepsilon} \tag{4}$$

where the matrix $\mathbf{W} = \mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}$ given in (4) is generally known as weight matrix or leverage matrix. The weight matrix **W** reflects joint effect of *k* regressors on the fitted responses. Writing the data matrix of *k* explanatory variables as $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]^\mathbf{T}$, the *i*-th diagonal element of the weight matrix **W** is defined as

$$w_{ii} = \mathbf{x_i}^\mathbf{T}(\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{x_i} \tag{5}$$

For a perfect balanced design, w_{ii} can be written as

$$w_i = \frac{1}{n} + \frac{(x_{i1} - \bar{x}_{.1})^2}{\sum (x_{i1} - \bar{x}_{.1})^2} + \frac{(x_{i2} - \bar{x}_{.2})^2}{\sum (x_{i2} - \bar{x}_{.2})^2} + ... + \frac{(x_{ip} - \bar{x}_{.p})^2}{\sum (x_{ip} - \bar{x}_{.p})^2}$$

and thus the diagonal elements w_{ii} of the weight matrix **W** are considered as leverage values, which measure influence of each observation in the *X*-space.

A good number of works have been done in the detection of a single high leverage point. It is easy to show that the average value of w_{ii} is k/n, where *k* is the number of the regressors (including the intercept term) and *n* is the total number of observations. Data points having large w_{ii} values are generally considered as high leverage points. Since finding the theoretical distribution of w_{ii} is difficult, all of the high leverage detection techniques are based on rules of thumb. Hoaglin and Welsch (1978) considered observations to be unusual when w_{ii} exceeds $2k/n$ which is also known as the twice-the-mean (2M) rule. Vellman and Welsch (1981) preferred the thrice-the–mean (3M) rule where w_{ii} is considered to be large when it exceeds $3k/n$. Huber (1981) suggested breaking the range of possible values, $(0 \le w_{ii} \le 1)$ into three intervals. Values $w_{ii} \le 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and values above 0.5 should be avoided. Well known Mahalanobis distances are also suggested to use as measures of leverages in the literature, however, Rousseeuw and Leroy (1987) showed that Mahalanobis distance for each of the points has a one-one relationship with w_{ii} and do not yield any extra information in the leverage structure of a data point. Hadi (1992) pointed out that traditionally used measures of leverages are not sensitive enough to the high leverage points. He introduced a single case deleted leverage measure, named as potential, which is believed to be more sensitive to the high leverage point. Imon and Khan (2003b) showed that in the presence of multiple high leverage points, observations are masked in such a way that even potential values may not focus on all of them. As a remedy to this problem, Imon (2002) proposed generalized potentials for the identification of multiple high leverage points in linear regression. Further developments of the generalized potentials are done by Habshah et al. (2009) and Bagheri et al. (2009).

3. Estimation of the Fixed-*X* in LFRM

All the leverage measures discussed in section 2 are designed for fixed-X model and hence cannot be readily applied to errors in variable models. In this section we obtain the estimated values of *X* so that these values can be used as fixed-*X* in the subsequent studies.

Let us assume,

$$E(\delta_i) = E(\varepsilon_i) = 0, \text{var}(\delta_i) = \sigma_\delta^2, \text{var}(\varepsilon_i) = \sigma_\varepsilon^2, \forall i$$
$$\text{cov}(\delta_i, \delta_j) = \text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$$
$$\text{cov}(\delta_i, \varepsilon_j) = 0, \forall i, j \quad (6)$$

Model (2) is also known as the unreplicated linear functional relationship when there is only one relationship between the two variables X and Y. There are $(n+4)$ parameters to be estimated, which are $\beta, \alpha, \sigma^2, \tau^2$ and $X_1, X_2, ..., X_n$. Several methods of parameter estimation have been developed (Fuller, 1987) but our primary interest is the maximum likelihood (ML) method. Let (2) and (6) hold, and that δ_i and ε_i are independent normal variables, viz.

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2) \text{ and } \delta_i \sim N(0, \sigma_\delta^2) \quad (7)$$

Since X_i are non-random variables, $\sigma_x^2 = 0$ and there are $(n+4)$ parameters, namely $\beta, \alpha, \sigma^2, \tau^2$ and the n values of X_i to be estimated. The model (2) yields the 2-vector $(X_i, Y_i)^T$, $i = 1,2,...,n$ distributed as

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} \sim N\left(\begin{bmatrix} X_i \\ Y_i \end{bmatrix}, \begin{bmatrix} \sigma_\delta^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{bmatrix} \right) \quad (8)$$

and the likelihood function is given by

$$L(\beta, \alpha, \sigma_\delta^2, \sigma_\varepsilon^2, X_1, X_2, ..., X_n) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma_\delta\sigma_\varepsilon} \exp - \left[\frac{(x_i - X_i)^2}{2\sigma_\delta^2} + \frac{(y_i - \alpha - \beta X_i)^2}{2\sigma_\varepsilon^2} \right]$$
$$= \frac{1}{(2\pi\sigma_\delta\sigma_\varepsilon)^n} \exp - \left[\frac{\sum (x_i - X_i)^2}{2\sigma_\delta^2} + \frac{\sum (y_i - \alpha - \beta X_i)^2}{2\sigma_\varepsilon^2} \right] \quad (9)$$

If the ratio of the error variances $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ is known and taking log for (9) becomes

$$\log L = -n\log 2\pi - \frac{n}{2}(2\log\sigma_\varepsilon^2 - \log\lambda) - \frac{1}{2\sigma_\varepsilon^2}\left[\lambda\sum(x_i - X_i)^2 + \sum(y_i - \alpha - \beta X_i)^2\right] \quad (10)$$

Now differentiating (10) with respect to parameters β, α, σ^2 and X_i, we proceed to the ML solution

$$\frac{\partial \log L}{\partial \alpha} = \frac{1}{\hat\sigma_\varepsilon^2}\sum(y_i - \hat\alpha - \hat\beta\hat X_i) = 0 \quad (11)$$

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\hat\sigma_\varepsilon^2}\sum\hat X_i(y_i - \hat\alpha - \hat\beta\hat X_i) = 0 \quad (12)$$

$$\frac{\partial \log L}{\partial X_i} = \frac{\lambda}{\hat\sigma_\varepsilon^2}(x_i - \hat X_i) + \frac{1}{\hat\sigma_\varepsilon^2}\hat\beta(y_i - \hat\alpha - \hat\beta\hat X_i) = 0 \quad (13)$$

$$\frac{\partial \log L}{\partial \sigma_\varepsilon} = -\frac{2n}{\hat\sigma_\varepsilon} + \frac{1}{\hat\sigma_\varepsilon^3}\left[\lambda\sum(x_i - \hat X_i)^2 + \sum(y_i - \hat\alpha - \hat\beta\hat X_i)^2\right] = 0 \quad (14)$$

From (14) the estimator of σ_ε^2 is derived as

$$\hat\sigma_\varepsilon^2 = \frac{1}{2n}\left[\lambda\sum(x_i - \hat X_i)^2 + \sum(y_i - \hat\alpha - \hat\beta\hat X_i)^2\right] \quad (15)$$

which is not consistent. Kendall and Stuart (1979) showed a consistent estimator of σ_ε^2 can be derived by multiplying $\frac{2n}{n-2}$ to (15), that is

$$\hat\sigma_\varepsilon^2 = \frac{1}{n-2}\left[\lambda\sum(x_i - \hat X_i)^2 + \sum(y_i - \hat\alpha - \hat\beta\hat X_i)^2\right] \quad (16)$$

Using (11) and (13), we obtain the estimated values of *X* as

$$\hat X_i = \frac{\lambda x_i + \hat\beta(y_i - \hat\alpha)}{(\lambda + \hat\beta^2)} \quad (17)$$

where $$\hat\alpha = \bar y - \hat\beta\bar x \quad (18)$$

and

$$\hat\beta = \frac{\left(\sum\hat X_i y_i - \hat\alpha\sum\hat X_i\right)}{\sum\hat X_i{}^2} = \frac{(\lambda+\hat\beta^2)(\lambda S_{xy} + \hat\beta S_{yy} + n\hat\beta^3\bar x^2 + n\lambda\hat\beta\bar x^2)}{\lambda^2\sum y_i{}^2 + 2\lambda\hat\beta S_{xy} + \hat\beta^2 S_{yy} + n\hat\beta^4\bar x^2 + 2n\lambda\hat\beta^2\bar x^2}$$

gives $S_{xy}\hat\beta^2 + (\lambda S_{xx} - S_{yy})\hat\beta - \lambda S_{xy} = 0$ and that implies

$$\hat\beta = \frac{(S_{yy} - \lambda S_{xx}) + \sqrt{(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2}}{2S_{xy}} \quad (19)$$

where, $\bar y = \frac{\sum y_i}{n}, \bar x = \frac{\sum x_i}{n}, S_{yy} = \sum y_i{}^2 - n\bar y^2, S_{xx} = \sum x_i{}^2 - n\bar x^2$ and $S_{xy} = \sum x_i y_i - n\bar x\bar y$.

4. Identification of High Leverage Points in LFRM

In this section we suggest a procedure for the identification of high leverage points in linear functional relation model. From a set of observed x and y (both assumed to be measured with error), we have estimated the fixed-*X* using the equation (17), i.e., using

$$\hat{X}_i = \frac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{(\lambda + \hat{\beta}^2)}$$

Since we have a single explanatory variable here, the formula (5) for the computation of leverage values can be simplified as

$$\hat{w}_{ii} = \hat{x}_i^T (\hat{X}^T \hat{X})^{-1} \hat{x}_i = \frac{1}{n} + \frac{\left(\hat{X}_i - \bar{\hat{X}}\right)^2}{\sum_{i=1}^{n} \left(\hat{X}_i - \bar{\hat{X}}\right)^2} \tag{20}$$

Since the above formula contains mean and sum of squares of means which could be very sensitive to high leverage points. For this reason we propose a new formula for the leverages analogous to formula (20), but here the non-robust components are replaced by their corresponding robust alternatives. Hence the formula is

$$\tilde{w}_{ii} = \frac{1}{n} + \frac{|\hat{X}_i - Med\left(\hat{X}_i\right)|}{nMAD\left(\hat{X}_i\right)} \tag{21}$$

It is easy to show from (20) and (21) that

$$\text{mean}(\hat{w}_{ii}) = \text{median}(\tilde{w}_{ii}) = 2/n.$$

We consider several measures of the identification of high leverage points, the twice the mean (2M) rule, the thrice-the-mean (3M) rule, and then introduce a new cut-off point. Since it may not be easy to find the theoretical distribution of $\tilde{w}_{ii}$ and often excessive high leverage values can affect measures like mean and standard deviation, we define a confidence bound type cut-off point

$$\tilde{w}_{ii} > \text{Median}(\tilde{w}_{ii}) + 3\ \text{MAD}(\tilde{w}_{ii}) \tag{22}$$

which is analogous to forms used by Hadi (1992), Imon (2002,2005) and others.
In this paper, we consider five identification rules which are listed below:

Rule 1 (Classical 2M): $\hat{w}_{ii} > 4/n$

Rule 2 (Classical 3M): $\hat{w}_{ii} > 6/n$

Rule 3 (New 2M based on Median): $\tilde{w}_{ii} > 4/n$

Rule 4 (New 3M based on Median): $\tilde{w}_{ii} > 6/n$

Rule 5 (New Median based Cut-off point): $\tilde{w}_{ii} > \text{Median}(\tilde{w}_{ii}) + 3\ \text{MAD}(\tilde{w}_{ii})$

We compare the performances of the above rules in terms of correct identification of high leverage points and swamping rate of good leverages.

5. Monte Carlo Simulations

In this section we report a Monte Carlo simulation which is designed to investigate the performances of different measures of leverages in linear functional relation model. For four different sample sizes, n = 20, 30, 50 and 100, we generated the X values from Uniform (20, 40). Here we consider three different percentages, i.e., 10%, 20%, and 30% high leverage points. The X value corresponding to the lowest high leverage value is then set at 100 and the next values have an increment of 5 each. To generate a model like (2), we then define $x_i = X_i + \delta_i$, where δ_i is N (0, 1). The values of y_i are generated as

$$y_i = 20 + 2X_i + \varepsilon_i$$

where ε_i is also N (0, 1). For each different sample we apply all five leverage identification rules mentioned in section 4 and compute the correct identification rate (IR) and the swamping rate (SR) in terms of percentages. We run 10,000 simulations for each combination and these results are presented in Table 1. When no high leverage point exists, we observe from Table 1 that for n = 20, all methods considered in the simulation perform well. However, rule 1, i.e., the traditional leverage measure based on the 2M rule has about 5% swamping rate. The newly proposed rule 4 performs the best as its swamping rate is the lowest followed by rule 2, rule 5 and rule 3. The performance of all these rules tend to improve with the increase in sample sizes but still rule 1 has relatively very high swamping rate which clearly shows that the 2M rule is too prone to declare low leverage points as points of high leverages. In

case of 10% high leverages, almost all methods perform very well. Each method maintains 100% identification rate with low swamping rate. Only when n = 100, the identification rate for rule 2 is 90%. But when 20% or 30% high leverage points are present in the data both the 2M and the 3M rule break down. The rule 2, i.e., the 3M performs worst as often its correct identification rate is 0%. The performance of the rule 1 is also poor as it can identify around 13% cases correctly when there is 30% contamination. The performances of the newly proposed all three rules, i.e., rules 3, 4 and 5 are very satisfactory. They have almost 100% correct identification rate with very small swamping rates, if at all.

Table 1. Identification and Swamping Rates of Different Leverage Detection Rules

Sample Size	Rules	Percentages of High Leverage Points							
		0		10		20		30	
		IR(%)	SR(%)	IR(%)	SR(%)	IR(%)	SR(%)	IR(%)	SR(%)
n = 20	Rule 1	-----	4.4985	**100.00**	0.0000	84.01	0.0000	16.67	0.0000
	Rule 2	-----	0.2315	**100.00**	0.0000	00.00	0.0000	0.0000	0.0000
	Rule 3	-----	2.4510	**100.00**	1.1551	**100.00**	0.8081	**100.00**	0.1350
	Rule 4	-----	0.1355	**100.00**	0.0689	**100.00**	0.0306	**99.45**	0.0043
	Rule 5	-----	0.8765	**100.00**	0.3372	**100.00**	0.0712	**100.00**	0.0057
n = 30	Rule 1	-----	4.1257	**100.00**	0.0000	82.95	0.0000	32.27	0.0000
	Rule 2	-----	0.1390	**100.00**	0.0000	16.67	0.0000	0.00	0.0000
	Rule 3	-----	1.2467	**100.00**	0.7541	**100.00**	0.3775	**100.00**	0.0362
	Rule 4	-----	0.0163	**100.00**	0.0111	**100.00**	0.0029	**99.82**	0.0000
	Rule 5	-----	0.1137	**100.00**	0.1137	**100.00**	0.0150	**99.90**	0.0005
n = 50	Rule 1	-----	3.893	**100.00**	0.0000	70.00	0.0000	33.33	0.0000
	Rule 2	-----	0.0544	**100.00**	0.0000	30.00	0.0000	0.00	0.0000
	Rule 3	-----	0.4366	**100.00**	0.2300	**100.00**	0.1238	**100.00**	0.0046
	Rule 4	-----	0.0006	**100.00**	0.0002	**100.00**	0.0000	**99.98**	0.0000
	Rule 5	-----	0.0740	**100.00**	0.0178	**100.00**	0.0008	**99.99**	0.0000
n = 100	Rule 1	-----	3.5172	**100.00**	0.0000	60.00	0.0000	36.67	0.0000
	Rule 2	-----	0.0152	90.00	0.0000	35.00	0.0000	13.33	0.0000
	Rule 3	-----	0.0543	**100.00**	0.0327	**100.00**	0.0136	**100.00**	0.0004
	Rule 4	-----	0.0000	**100.00**	0.0000	**100.00**	0.0000	**100.00**	0.0000
	Rule 5	-----	0.0050	**100.00**	0.0044	**100.00**	0.0000	**100.00**	0.0000

6. Example

We consider a real world data to investigate the performance of our proposed method. In order to make the relationship as model (2), we assume that measurement error can occur in either variable of this example.

6.1. Iron in Slag Data

This example is taken from Hand et al. (1994) where the data for 50 results of iron content of crushed blast furnace slag measured by two different techniques, which are chemical test (Y) and magnetic test (X). The original data together with the estimated X values by the maximum likelihood formula (17) is presented in Table 2. Now we compute the leverage values for this data set and these values are presented in Table 2. Here the cut-off point for rule 1 and 3 is 0.08, for rule 2 and 4 is 0.12 and rule 5 is 0.0975 respectively. We observe from the Table 3 that the traditional leverage values $\hat{w}_{ii}$ do not identify any high leverage points, but the 2M rule swamps in six good cases. The newly proposed leverage measures $\tilde{w}_{ii}$ do not identify any high leverage points but the 2M rule swamps in one good case. The 3M rule does not identify any high leverage point for both of these two leverage measures. We observe exactly the same performance from the rule based on the new cut-off point as well.

Table 2: Iron in Slag Data

Index	Chemical Test (y)	Magnetic Test (x)	Estimated X	Index	Chemical Test (y)	Magnetic Test (x)	Estimated X
1	24	25	24.39586	26	15	15	14.50594
2	20	21	20.24055	27	25	16	20.03000

3	16	22	18.78758	28	16	16	15.54476
4	20	21	20.24055	29	15	16	15.04640
5	24	17	20.07211	30	15	16	15.04640
6	25	21	22.73235	31	16	16	15.54476
7	18	21	19.24383	32	17	12	13.88125
8	27	25	25.89094	33	27	26	26.43141
9	18	20	18.70336	34	19	15	16.49937
10	22	22	21.77774	35	27	28	27.51235
11	10	13	10.93320	36	16	15	15.00430
12	20	18	18.61914	37	30	28	29.00743
13	14	16	14.54805	38	15	15	14.50594
14	24	21	22.23399	39	29	30	29.59001
15	16	14	14.46383	40	15	15	14.50594
16	24	18	20.61258	41	26	32	29.17587
17	18	19	18.16289	42	13	17	14.59015
18	23	20	21.19516	43	25	28	26.51563
19	29	25	26.88766	44	24	18	20.61258
20	21	23	21.81985	45	22	16	18.53492
21	27	20	23.18860	46	21	18	19.11750
22	20	20	19.70008	47	28	33	30.71306
23	23	18	20.11422	48	24	22	22.77446
24	21	19	19.65797	49	25	33	29.21798
25	19	19	18.66125	50	15	20	17.20828

6.2. Modified Iron in Slag Data

Next we modified the original iron in slag data by inserting few high leverage points. We consider three different situations. In case 1, 5 low leverage cases (10%) are replaced by high leverage points. In case 2 and case 3 we replace 20% and 30% low leverage points by points of high leverages respectively. Table 4 gives the first 15 observations of the modified iron in slag data and the corresponding leverage values are given in Table 5. Now we compute the leverage values for this data set and these values are presented in Table 5. Here the cut-off point for rule 1 and 3 is 0.08 and for rule 2 and 4 is 0.12 as they were before. The cut-off points for rule 5 are 0.1052, 0.1024, and 0.0845 for 10%, 20% and 30% high leverage points respectively. We observe from the Table 5 that for the 10% contamination, the traditional leverage values $\hat{w}_{ii}$ can identify high leverage points successfully, but their performances tend to deteriorate with the increase in the level of contamination. For 20% contamination it fails to identify four high leverage cases out of 10 and for the 30% contamination it fails to identify 10 out of 15 high leverage points. The newly proposed leverage measures $\tilde{w}_{ii}$ perform very well in this regard. All high leverage points are successfully identified irrespective of the level of contamination.

Table 3: Leverage Values for the Iron in Slag Data

Index	$\hat{w}_{ii}$	$\tilde{w}_{ii}$	Index	$\hat{w}_{ii}$	$\tilde{w}_{ii}$
1	0.0343	0.0470	26	0.0482	0.0520
2	0.0200	0.0222	27	0.0200	0.0209
3	0.0218	0.0264	28	0.0389	0.0458
4	0.0200	0.0222	29	0.0431	0.0488
5	0.0200	0.0212	30	0.0431	0.0488
6	0.0250	0.0371	31	0.0389	0.0458
7	0.0209	0.0237	32	0.0546	0.0557
8	0.0466	0.0560	33	0.0520	0.0592
9	0.0221	0.0269	34	0.0320	0.0401
10	0.0218	0.0314	35	0.0642	0.0657
11	0.0939	0.0734	36	0.0435	0.0490

12	0.0223	0.0274	37	**0.0844**	0.0746
13	0.0478	0.0517	38	0.0482	0.0520
14	0.0232	0.0341	39	**0.0933**	0.0781
15	0.0486	0.0522	40	**0.0482**	0.0520
16	0.0200	0.0244	41	**0.0869**	0.0756
17	0.0237	0.0301	42	0.0474	0.0515
18	0.0207	0.0279	43	0.0529	0.0597
19	0.0569	0.0619	44	0.0200	0.0244
20	0.0220	0.0316	45	0.0225	0.0279
21	0.0271	0.0398	46	0.0211	0.0244
22	0.0202	0.0209	47	**0.1121**	**0.0848**
23	0.0200	0.0214	48	0.0252	0.0373
24	0.0203	0.0212	49	**0.0876**	0.0759
25	0.0222	0.0271	50	0.0279	0.0358

Table 4: The First 15 Observations of the Modified Iron in Slag Data

Index	(y)	(x) with 10% HLP	(x) with 20% HLP	x with 30% HLP
1	24	25 **(50)**	25 **(50)**	25 **(50)**
2	20	21 **(55)**	21 **(55)**	21 **(55)**
3	16	22 **(60)**	22 **(60)**	22 **(60)**
4	20	21 **(65)**	21 **(65)**	21 **(65)**
5	24	17 **(70)**	17 **(70)**	17 **(70)**
6	25	21	21 **(75)**	21 **(75)**
7	18	21	21 **(80)**	21 **(80)**
8	27	25	25 **(85)**	25 **(85)**
9	18	20	20 **(90)**	20 **(90)**
10	22	22	22 **(95)**	22 **(95)**
11	10	13	13	13 **(100)**
12	20	18	18	18 **(105)**
13	14	16	16	16 **(110)**
14	24	21	21	21 **(115)**
15	16	14	14	14 **(120)**

Table 5: Leverage Values for the Modified Iron in Slag Data

Index	$\hat{w}_{ii}$			$\tilde{w}_{ii}$		
	10% HLP	20% HLP	30% HLP	10% HLP	20% HLP	30% HLP
1	**0.0936**	0.0349	0.0219	**0.1612**	**0.1376**	**0.0985**
2	**0.1249**	0.0435	0.0244	**0.1842**	**0.1571**	**0.1128**
3	**0.1618**	0.0541	0.0278	**0.2072**	**0.1766**	**0.1271**
4	**0.2042**	0.0666	0.0322	**0.2303**	**0.1960**	**0.1414**
5	**0.2522**	**0.0811**	0.0375	**0.2533**	**0.2155**	**0.1557**
6	0.0205	**0.0976**	0.0438	0.0319	**0.2350**	**0.1700**
7	0.0215	**0.1160**	0.0511	0.0253	**0.2544**	**0.1843**
8	0.0204	**0.1364**	0.0593	0.0518	**0.2739**	**0.1985**
9	0.0224	**0.1587**	0.0685	0.0208	**0.2934**	**0.2128**
10	0.0204	**0.1830**	0.0786	0.0336	**0.3128**	**0.2271**
11	0.0393	0.0330	**0.0897**	0.0582	0.0494	**0.2414**
12	0.0243	0.0261	**0.1018**	0.0262	0.0270	**0.2557**
13	0.0299	0.0287	**0.1149**	0.0409	0.0365	**0.2700**
14	0.0207	0.0233	**0.1289**	0.0309	0.0258	**0.2843**
15	0.0334	0.0310	**0.1439**	0.0480	0.0437	**0.2986**
16	0.0233	0.0258	0.0291	0.0225	0.0257	0.0328
17	0.0236	0.0253	0.0283	0.0236	0.0237	0.0299

18	0.0215	0.0241	0.0275	0.0255	0.0216	0.0271
19	0.0206	0.0209	0.0242	0.0537	0.0428	0.0271
20	0.0201	0.0221	0.0254	0.0371	0.0326	0.0214
21	0.0209	0.0239	0.0275	0.0293	0.0228	0.0271
22	0.0220	0.0243	0.0275	0.0227	0.0207	0.0271
23	0.0235	0.0259	0.0291	0.0234	0.0260	0.0328
24	0.0228	0.0251	0.0283	0.0208	0.0228	0.0299
25	0.0233	0.0252	0.0283	0.0227	0.0234	0.0299
26	0.0316	0.0299	0.0318	0.0445	0.0401	0.0414
27	0.0257	0.0278	0.0309	0.0305	0.0332	0.0385
28	0.0290	0.0286	0.0309	0.0390	0.0359	0.0385
29	0.0295	0.0287	0.0309	0.0400	0.0362	0.0385
30	0.0295	0.0287	0.0309	0.0400	0.0362	0.0385
31	0.0290	0.0286	0.0309	0.0390	0.0359	0.0385
32	0.0380	0.0337	0.0349	0.0561	0.0511	0.0499
33	0.0209	0.0206	0.0236	0.0563	0.0461	0.0299
34	0.0298	0.0295	0.0318	0.0407	0.0389	0.0414
35	0.0226	0.0201	0.0227	0.0653	0.0539	0.0357
36	0.0311	0.0298	0.0318	0.0435	0.0398	0.0414
37	0.0234	0.0201	0.0227	0.0681	0.0548	0.0357
38	0.0316	0.0299	0.0318	0.0445	0.0401	0.0414
39	0.0259	0.0200	0.0218	0.0762	0.0622	0.0414
40	0.0316	0.0299	0.0318	0.0445	0.0401	0.0414
41	0.0282	0.0201	0.0211	0.0824	0.0691	0.0471
42	0.0283	0.0277	0.0300	0.0373	0.0329	0.0357
43	0.0222	0.0201	0.0227	0.0634	0.0532	0.0357
44	0.0233	0.0258	0.0291	0.0225	0.0257	0.0328
45	0.0267	0.0280	0.0309	0.0334	0.0341	0.0385
46	0.0240	0.0260	0.0291	0.0253	0.0267	0.0328
47	0.0311	0.0203	0.0209	0.0888	0.0735	0.0499
48	0.0202	0.0226	0.0261	0.0354	0.0297	0.0214
49	0.0298	0.0203	0.0209	0.0859	0.0726	0.0500
50	0.0231	0.0246	0.0275	0.0219	0.0207	0.0271

7. Conclusions

In this paper, our main objective is to propose a new method for the identification of high leverage points in linear functional relationship model. After obtaining a method of finding the fixed-X values, we propose three different identification rules based on robust measures of leverages. Both numerical and simulation results show that the traditionally used measures may often fail to identify even a single high leverage point when 20% to 30% high leverage points are present in the data. The 2M rule based on traditional leverage measure possesses relatively very high swamping rate as well. However, the proposed methods perform very well in every occasion. Our study clearly shows that they can correctly identify all high leverage points without swamping low leverage cases.

Acknowledgement
The authors are thankful to the Referees and Editor of PJSOR for their very helpful comments and suggestions.

References

1. Abdullah, M. B. (1995). Detection of influential observations in functional errors-in- variables model. Communications in Statistics: Theory and Methods. 24:1585–1595.
2. Bagheri, A., Habshah, M. and Imon, A.H.M.R. (2009). Two-step robust diagnostic method for identification of multiple high leverage points. Journal of Mathematics and Statistics. 5: 97–206.
3. Chatterjee, S. and Hadi, A. S. (1988). Sensitivity Analysis in Linear Regression, Wiley, New York.
4. Fuller, W.A. (1987). *Measurement error models*, Wiley, New York.

5. Habshah, M., Norazan, R. and Imon, A.H.M.R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. Journal of Applied Statistics. 36: 507–520.
6. Hadi, A.S. (1992). A new measure of overall potential influence in linear regression. Computational Statistics and Data Analysis. 14: 1-27.
7. Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994) A Handbook of Small Data Sets, Chapman and Hall, London.
8. Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and ANOVA. The American Statistician. 32:17-22.
9. Hocking, R.R. and Pendleton, O.J. (1983). The regression dilemma. Communications in Statistics-Theory and Methods. 12: 497-527.
10. Huber, P.J. (1981). Robust Statistics. Wiley, New York.
11. Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression. Journal of Statistical Studies. 3(Special Volume): 207–218.
12. Imon, A.H.M.R. (2005). Identifying multiple influential observations in linear regression. Journal of Applied Statistics. 32: 929 – 946.
13. Imon, A. H. M. R. (2009). Deletion residuals in the detection of heterogeneity of variances in linear regression. Journal of Applied Statistics. 36:347–358.
14. Imon, A. H. M. R. and Khan, M.A.I. (2003a). A solution to the problem of multicollinearity caused by the presence of multiple high leverage points. International Journal of Statistical Sciences. 2:37–50.
15. Imon, A.H.M.R. and Khan, M.A.I. (2003b). A comparative study on the identification of high leverage points in linear regression. Journal of Statistical Studies. 23: 27–32.
16. Kamruzzaman, M. and Imon, A. H. M. R. (2002). High leverage point: Another source of multicollinearity. Pakistan Journal of Statistics.18: 435–448.
17. Kendall, M.G. and Stuart, A. (1979). The Advance Theory of Statistics, Vol.2, Griffin, London.
18. Mahdizadeh, M., and Zamanzade, E. (2020). Estimating asymptotic variance of M-estimators in ranked set sampling. Computational Statistics. https://doi.org/10.1007/s00180-019-00946-3
19. Mahdizadeh, M., and Zamanzade, E. (2019). Efficient body fat estimation using multistage pair ranked set sampling. Statistical Methods in Medical Research. 28 (1): 223-234.
20. Mahdizadeh, M. and Zamanzade, E. (2018). Smooth estimation of a reliability function in ranked set sampling, *Statistics*: A Journal of Theoretical and Applied Statistics. 52(4): 750-768.

21. Peña, D.and Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. Journal of the Royal Statistical Society Ser- B. 11(57): 18-44.
22. Ryan, T.P. (1997). Modern Regression Methods, Wiley, New York.
23. Rousseeuw, P.J. and Leroy, A. (1987). Robust Regression and Outlier Detection, Wiley, New York.
24. Vellman, P.F. and Welsch, R.E. (1981). Efficient computing of regression diagnostics. The American Statistician. 35:234-42.
25. Vidal, I., Iglesias, P. and Galea, M. (2007). Influential observations in the functional measurement error model. Journal of Applied Statistics. 34:1165-83.
26. Wellman, M. J. and Gunst, R. F. (1991). Influence diagnostics for linear measurement error models. Biometrika. 78: 373–380.
27. Zamanzade, E., and Mahdizadeh, M. (2020). Using ranked set sampling with extreme ranks in estimating the population proportion. Statistical Methods in Medical Research. 29 (1): 165-177.