World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:3, No:12, 2009

On the outlier Detection in Nonlinear Regression

A. Hossein Riazoshams, B. Midi Habshah, Jr., C. Mohamad Bakri Adam

***Abstract*—**The detection of outliers is very essential because of their responsibility for producing huge interpretative problem in linear as well as in nonlinear regression analysis. Much work has been accomplished on the identification of outlier in linear regression, but not in nonlinear regression. In this article we propose several outlier detection techniques for nonlinear regression. The main idea is to use the linear approximation of a nonlinear model and consider the gradient as the design matrix. Subsequently, the detection techniques are formulated. Six detection measures are developed that combined with three estimation techniques such as the Least-Squares, M and MM-estimators. The study shows that among the six measures, only the studentized residual and Cook Distance which combined with the MM estimator, consistently capable of identifying the correct outliers.

***Keywords*—**Nonlinear Regression, outliers, Gradient, Least Square, M-estimate, MM-estimate.

I. Introduction

MANY statistics practitioners have been using residuals for the identification of outliers. The use of residuals resulting from the ordinary least squares (OLS) estimates will give a misleading conclusion because the residuals are functions of leverages and true errors. According to Habshah et al. [9], the high leverage points together with large errors (outliers) and the residuals are responsible for the cause of masking and swamping of outliers in linear regressions. There are a considerable amount of good written papers relating to identification of outliers in linear regression [see for example, Hadi [10], Habshah et al. [9], Cook and Weisberg [7], Belsley et al.[6], Anscombe and Tukey [1] and the discussion on the properties of Atkinson's distances in [3] and [4]). However, not much work has been explored in the formulation of the outlier's identification method in nonlinear regression. Cook and Weisberg [6] and Fox et al. [7] introduced a measure for the identification of outliers in nonlinear model, which is based on the OLS method. However, it is now evident that outliers have an adverse effect on the OLS estimates (see for example Habshah et al. [9]). In this situation, we suspect that any measures which are based on the OLS estimates are not efficient and this may cause swamping (false positive) and masking (false negative) effects. In this paper, an attempt is made to propose robust method of identification of outliers in nonlinear model.

F. A. Correspondence Author, Institute for Mathematical Research, University Putra Malaysia, Lab of Applied and Computational Statistics, 43400 Serdang Malaysia (corresponding phone: +60176518634; e-mail: riazi_hosein2@yahoo.com.sg).

S. B. Institute for Mathematical Research, University Putra Malaysia, Lab of Applied and Computational Statistics 43400 Serdang Malaysia (e-mail: robust.statistics@gmail.com).

T. C. Institute for Mathematical Research, University Putra Malaysia, Lab of Applied and Computational Statistics, 43400 Serdang Malaysia (e-mail: bakri@math.upm.edu.my).

II. Robust Nonlinear Regression

Consider the nonlinear model with additive error terms:

$$\equiv \eta(\theta) + \tag{1}$$

where $\mathbf{y} = [y_1, y_2, ..., y_n]^T$ is $n \times 1$ response vector, $\eta(\theta) = \left[f(\mathbf{x}_1; \theta), ..., f(\mathbf{x}_n; \theta) \right]$ is $n \times 1$ vector of function models $f(x_i; \theta)$'s, $\mathbf{x}_i = [x_{i1}, x_{i2}, ..., x_{ik}]^T$ is k dimensional predictor (design) vector, $= [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]^T$ is $n \times 1$ vector of iid residuals, which under normality assumption assumed to have normal distribution with mean zero and variance $\sigma^2 \mathrm{I}_n$, and $\theta \in \Re^p$ and p dimensional unknown parameter vector. The least squares estimator, θ of the nonlinear regression in (3) are found by minimizing the error sum of squares:

$$\hat{\theta}_{LS} = \arg\min_{\theta} \|\mathbf{r}\|^2 \tag{2}$$

where $\boldsymbol{r}$ is the residual vectors with elements $r_i = y_i - f(x_i; \theta)$, and ||.|| is the Euclidean norm. However, many statistics practitioners are not aware of the fact that outliers have an unduly effects on the OLS estimates. As an alternative, robust methods which are not easily affected by outliers are put forward to remedy these problems. There are many robust methods in the literatures and in this paper, only the M and MM estimators are considered (See Huber [14] and Stromberg [20]).

The M estimator is obtained by minimizing:

$$\hat{\theta}_M = \arg\min_{\theta} h(\theta) \tag{3}$$

where $h(\theta)$ is given by

$$h(\theta) = \sum_{i=1}^{n} \rho\left[\frac{r_i(\theta)}{\sigma} \right]$$

and $\rho(\theta)$ is a robust loss function satisfies the Huber conditions (see [14]). The Newton method (see [21]) is used to estimate the parameter theta. When convergence is not achieved due to large residuals, the Levenberg-Marquardt is utilized.

Yohai [22] and Stromberg [20] introduced the the MM estimator in linear and nonlinear regression, respectively. Stromberg proposed the computation of MM estimator in three stages as follows;

Stage 1 : Obtain a consistent high breakdown estimator

Stage 2 : Use stage 1 to calculate the M-estimate of scale using rho function ρ_0

Stage 3 : Compute the M estimate using rho function ρ_1 by using stage 1 and stage 2.

There are several rho functions, to choose from, and in this study, the Hampel redescending rho function (See [11]) denoted as ρ_H, is used in the analysis. Yohai (See [22]) revealed that $\rho_0(r)$ and $\rho_1(r)$ can be taken to be $\rho_H(r/k_0)$ and $\rho_H(r/k_1)$, respectively. Stromberg [20] demonstrated that selecting $k_0 = 0.212$ and $k_1 = .9014$ will guarantee a high breakdown estimate and will result in 95% efficiency under normal errors, respectively. The parameter estimates computed by these three techniques will be utilized in the development of the outlier measures in nonlinear regression that is discussed in Section III.

III. The Outlier Measures

Consider the multiple linear regressions,

$$Y = XB + \varepsilon$$

Where matrix X is the explanatory variable $n \times p$, Y is $n \times 1$ vector of response vector, ε is identically independent distributed error vector, n is number of observations, β is p dimensional unknown vector of coefficients. After the least squares estimates of the parameters β have been computed, the predicted value of the response variable can be written in the form of the Hat matrix as follows;

$$\hat{y} = X\hat{\beta} = Wy \tag{4}$$

where W is the hat matrix of

$$W = X(X^T X)^{-1} X^T \tag{5}$$

The elements of W are shown by w_{ij}. It can be seen from equation (4) that the influence of the response values on the prediction, depends on the values of w_{ij}. Equation (4) can be rewritten as:

$$\hat{y}_i = w_{ii} y_i + \sum_{j \neq i; j=1}^{n} w_{ij} y_j$$

Hoaglin and Welsh [13] suggested a direct use of w_{ij} as diagnostic of identifying high leverage points, if w_{ii} is large relative to the remaining terms. The fitted value $\hat{y}_i$ is more dominated by response $w_{ii} y_i$, so w_{ij} is interpreted as the amount of influence or leverage of $\hat{y}_j$ on y_i. In nonlinear regression, the linear approximation of function model is used, and replaces the explanatory matrix in linear regression, by the gradient of the function model. The linear approximation form can be derived by expanding the function model (1) around the true value θ^*

$$\eta(\theta) \cong \eta(\theta^*) + \dot{V}(\theta - \theta^*)$$

where $\dot{V} = \dfrac{\partial f(x;\theta)}{\partial \theta}$ is $n \times p$ gradient matrix computed at estimated point. Based on this approximation, an equivalent measure for equation (5) which is called as tangent plane leverage matrix is given by

$$H = \dot{V}(\dot{V}^T \dot{V})^{-1} \dot{V}^T \tag{6}$$

This leverage matrix in nonlinear plays a similar role as the Hat matrix W in linear form, as defined by equation (5) (see [7] p.187 and [16] Chapter 10).

Linear regression uses the Hat Matrix W as a beginning idea of influence detection tool, and creates several statistical measures for outlier detection. In this article, the leverage matrix H in Equation (6) is used in the formulation of the method of the identification of outliers in nonlinear regression.

In this section we extend the idea of influence outlier measures of linear regression for nonlinear case. Instead of using the Hat matrix W defined in (5), the gradient matrix H, as defined in (6) is utilized in the formulation of the influence measure.

A prevalent way of developing an influence detection method is to re fit a model with deleting a special case or a set of cases. Then, observe the amount of change of some statistics such as the parameter estimates, predicted, likelihoods, residuals, and so on, for a recalculated measure with the i'th data point, removed. The notation (-i) is used for each removed observation. It is important to point out that the three estimators namely the OLS, the M and the MM estimator are used to estimates the parameters of the nonlinear regression.

Subsequently, the respective estimates were utilized in the computation of the influence measures. The Six outlier measures are briefly discussed as follows;

A. Studentized and Deletion Studentized Residuals

This measure (hereafter refer as t_i) is used for identifying outliers. Suppose h_{ii} is the diagonal of leverage matrix H based on gradient in Equation (6), the studentized residual and the deleted studentized residuals are defined as follows

(See [19], [1]) :

$$t_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

and

$$d_i = \frac{r_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}}$$, respectively

where $\hat{\sigma}_{(-i)}$ is the estimated standard deviation in the absence of the *i'th* observation. The residuals, denoted as $r_i = y_i - f(x_i;\hat{\theta})$ is obtained from the OLS, M and the MM estimates.

The *i'th* observation is considered as an outlier if $|t_i| or |d_i| > 2.5 \text{ or } 3$.

B. Hadi potential

Hadi [10] proposed Hadi's potential denoted as pii to detect high leverage points or large residuals:

$$p_{ii} = \frac{h_{ii}}{1-h_{ii}}$$

Hadi [10] proposed a cut-off point for pii as $Median(p_{ii}) + c \cdot MAD(p_{ii})$ where MAD represents the Mean Absolute Deviance defined by:

$$MAD(p_{ii}) = Median\{p_{ii} - Median(p_{ii})\}/0.6745$$

C is an appropriately chosen constant such as 2 or 3.

C. Elliptic Norm (Cook Distance)

The Cook Distance (hereafter is refereed as CD) which is defined by Cook and Weisberg [7], is used to asses the influential observations. An observation is influence if the value of CD is greater than one. They defined CD as

$$CD_i(\dot{V}^T\dot{V}, p\hat{\sigma}^2) = (\theta - \hat{\theta}_{(-i)})^T(\dot{V}^T\dot{V})(\theta - \hat{\theta}_{(-i)}) \big/ p\hat{\sigma}^2$$

where $\hat{\theta}_{(-i)}$ is the parameter estimates when the *i'th* observation is removed. When $\hat{\theta}_{(-i)}$ is replaced by the linear approximation (see [7], and [8]), this norm changes to

$$CD_i(\dot{V}^T\dot{V}, p\hat{\sigma}^2) = \frac{t_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$

Where t_i and p is the studentized residual and the number of parameters in the model, respectively. With the cut of point equal to 1, that is the expectation of 50% confidence ellipsoid of parameter estimates.

D. Difference in Fits

Difference in Fits, denoted by DFFITS, is another diagnostics measure used in measuring the influence, defined by Belsley et al. [6]. For the *i'th* observation, DFFITS is defined as

$$DFFITS_i = \left(\sqrt{\frac{h_{ii}}{1-h_{ii}}}\right)|d_i|$$

where d_i is the deleted studentized residual. They considered observation is an outlier when DFFITS exceeds the cut of point equals to $2\sqrt{p/n}$.

E. Atkinson's Distance

Atkinson distance (hereafter refer as C_i) for observation i was developed by Atkinson [1] and it is used to detect the influential observation (See [3] and [4] for the discussion of the Atkinson's property). Atkinson [1] defined the Atkinson's distance as follows;

$$C_i = \left(\sqrt{\frac{n-p}{p}\frac{h_{ii}}{1-h_{ii}}}\right)|d_i|$$

where d_i is the deleted studentized residuals. He suggested a cut-off value equals to 2.

IV. Numerical example

In this section, a real data which is referred as the lake data, taken from Stromberg [20] is used to compare the preceding methods. The data set is collected from 29 lakes in Florida by United States Environmental Protection Agency (1978). Stromberg [20] has identified observations 10 and 23 as outliers.

The data presents the relationship between the mean annual total nitrogen concentration, TN, as the response variable and the average influence nitrogen concentration, NIN, and water retention time, TW, as predictors. The model associated with the data is:

$$TN_i = \frac{NIN_i}{1+\delta TW_i^{\beta}} + \varepsilon_i, \quad i = 1,...,29 \tag{7}$$

with unknown parameter vector $\theta = (\delta, \beta)$. The results of the six measures are shown in table 1 and 2.

The results of tables 1 and 2 suggest that most of the diagnostic measures that are based on the OLS and M estimates fail to identify the two observations as outliers. The p_{ii}-M based and p_{ii}-OLS based can correctly identify the two outliers but swamp two points (cases 11 and 16) as outliers. The results of Table 2 also point out that the d_i-MM based fail to identify any outlier. The p_{ii}-M based can correctly identify the two outliers but swamps one observation (case 16) as outliers. Furthermore, the DFFITS and C_i fail to detect observations 10 and 23 as outliers but swamp observation one as outlier. On the other hand, the studentized residuals and Cook's distance measures which are based on

the MM estimates identify correctly observations 10 and 23 as outliers.

TABLE I SIX OUTLIER MEASURES BASED ON OLS FOR LAKE DATA

Cut of points Index	t_i 3.0	d_i 3.0	CD_i 1.0	p_{ii} 0.066	$DFFITS_i$ 0.525	C_i 2.000
1	-1.525	-1.146	0.642	0.355	0.682	2.507
2	2.772	0.032	0.183	0.009	0.003	0.011
3	0.370	0.013	0.037	0.020	0.002	0.007
4	0.886	0.007	0.055	0.008	0.001	0.002
5	1.740	0.089	0.248	0.041	0.018	0.066
6	0.088	0.003	0.009	0.021	0.000	0.001
7	-0.860	-0.676	0.223	0.135	0.248	0.912
8	0.734	0.005	0.045	0.008	0.000	0.002
9	1.635	0.150	0.291	0.063	0.038	0.138
10	0.228	0.029	0.052	0.104	0.009	0.035
11	-1.259	-0.229	0.250	0.079	0.065	0.237
12	0.437	0.006	0.027	0.008	0.001	0.002
13	0.057	0.002	0.006	0.020	0.000	0.001
14	0.865	0.114	0.165	0.073	0.031	0.113
15	0.369	0.021	0.044	0.028	0.004	0.013
16	0.495	0.263	0.154	0.193	0.116	0.426
17	1.223	0.023	0.104	0.015	0.003	0.010
18	0.058	0.001	0.003	0.007	0.000	0.000
19	0.088	0.001	0.004	0.005	0.000	0.000
20	-0.380	-0.004	0.018	0.005	0.000	0.001
21	-0.007	0.000	0.001	0.016	0.000	0.000
22	1.240	0.008	0.064	0.005	0.001	0.002
23	-3.067	-11.428	4.528	4.359	23.861	87.673
24	1.458	0.038	0.127	0.015	0.005	0.017
25	-0.411	-0.010	0.031	0.011	0.001	0.004
26	-0.035	-0.001	0.003	0.018	0.000	0.001
27	0.137	0.019	0.027	0.079	0.005	0.020
28	-0.354	-0.012	0.032	0.016	0.001	0.005
29	0.264	0.002	0.014	0.006	0.000	0.001

V. SIMULATION STUDY

In this Section, a simulation study was performed to investigate whether the results of the simulation study confirm the conclusion of the real data set that the ti-MM based and CD-MM based are capable of identifying correct outliers.

The simulated value from the logistic model is based on the following function:

$$y_i = \frac{a}{1+b.e^{-cx_i}} + \varepsilon_i \qquad (8)$$

where $\theta = (a,b,c)$ are the parameters of the model. This model is chosen to mimic a real life chicken data set presented by Riazoshams and Midi (see [18]). In this simulation study, three parameters are considered (*a=2575, b=41, 0.11)* . The residuals are generated from normal distribution with mean zero and standard deviation $\sigma = 70$. The x_i's, are generated from a uniform distribution on [3,50] with a sample size 20.

Three different cases of contamination are considered,

Case A. The first good datum point (x_1, y_1), is replaced with a new value y_1 which is equals to $y_1 + 1000$.

Case B. The 6th, 7th and 8th data points are replaced with their corresponding y values increased by 1000 unit.

Case C. Six high leverage points were created by replacing the last six observations with (x,y) pair values (90,6500), (92,6510), (93,6400), (93,6520), (90,6600), (94,6600).

The six outlier detection techniques were then applied to the sets of A, B and C data based on the OLS, M and MM estimates. The results are exhibited in Tables 3-5. Due to space limitations, only the results on the outlier measures which are based on OLS and MM are presented. The results of the M based measures are discussed event though they are not displayed.

It can be observed from Table 3 that all methods fail to detect the single outlier except the t_i-OLS based, CD-M based t_i-MM based and CD-MM based. The results also point out that the p_{ii} based on the OLS, M and MM swamp two observations as outliers (cases 18 and 19).

It is interesting to note the results of Table 4, when there are three outliers in the y directions. The t_i-M based, the t_i-MM based and the CD-MM based are able to identify the 3 outliers correctly. Other outliers measures fail to identify even a single outlier. For instance, the p_{ii}-MM based masked the 3 outliers and swamps 2 observations (cases 18 and 19).

The presence of six high leverage points makes it harder for almost all outlier detection methods to detect high leverage points correctly. In this situation, most detection measures fail to identify even a single high leverage point because of the masking effects. It can be seen from Table 5 that again the ti-MM based and CD-MM based did a credible job. Both measures can identify the six high leverage points correctly.

TABLE II SIX OUTLIER MEASURES BASED ON M AND MM ESTIMATES FOR LAKE DATA

	M-estimate						MM-estimate					
Cut of points Index	t_i 3.0	d_i 3.0	CD_i 1.000	p_{ii} 0.067	$DFFITS_i$ 0.525	C_i 2.0	t_i 3.0	d_i 3.0	CD_i 1.0	p_{ii} 0.078	$DFFITS_i$ 0.525	C_i 2.0
1	-1.846	-1.945	0.799	0.375	1.191	4.374	-1.591	-1.780	0.469	0.174	0.742	2.728
2	3.823	0.041	0.248	0.008	0.004	0.014	1.505	2.477	0.159	0.022	0.371	1.362
3	0.583	0.023	0.054	0.017	0.003	0.011	0.035	0.000	0.002	0.008	0.000	0.000
4	1.244	0.011	0.077	0.008	0.001	0.004	-0.477	-0.063	0.030	0.008	0.006	0.020
5	2.497	0.091	0.334	0.036	0.017	0.063	1.091	0.010	0.110	0.020	0.001	0.005
6	0.204	0.008	0.020	0.019	0.001	0.004	-1.016	-0.014	0.077	0.012	0.002	0.006
7	-1.016	-1.249	0.264	0.135	0.459	1.688	-1.039	-1.086	0.182	0.061	0.269	0.988
8	1.035	0.008	0.063	0.007	0.001	0.002	-0.675	-0.078	0.042	0.008	0.007	0.025
9	2.365	0.142	0.390	0.054	0.033	0.122	1.689	0.009	0.182	0.023	0.001	0.005
10	0.373	0.052	0.083	0.100	0.017	0.061	-10.473	-0.132	5.615	0.575	0.100	0.367
11	-1.625	-0.681	0.352	0.094	0.209	0.767	-1.203	-1.054	0.260	0.093	0.322	1.183
12	0.648	0.012	0.037	0.007	0.001	0.004	0.299	0.000	0.012	0.003	0.000	0.000
13	0.157	0.011	0.015	0.018	0.001	0.005	-0.465	-0.001	0.030	0.009	0.000	0.000
14	1.312	0.163	0.241	0.067	0.042	0.155	1.020	0.010	0.119	0.027	0.002	0.006
15	0.587	0.040	0.067	0.026	0.006	0.024	0.373	0.003	0.027	0.011	0.000	0.001
16	0.873	0.498	0.291	0.223	0.235	0.863	1.326	0.145	0.379	0.163	0.058	0.215
17	1.746	0.034	0.141	0.013	0.004	0.014	0.828	-0.004	0.051	0.008	0.000	0.001
18	0.121	0.006	0.007	0.006	0.001	0.002	-0.017	0.001	0.001	0.003	0.000	0.000
19	0.158	0.005	0.007	0.004	0.000	0.001	-0.047	0.000	0.001	0.002	0.000	0.000
20	-0.498	-0.010	0.028	0.006	0.001	0.003	-0.076	0.023	0.012	0.049	0.005	0.019
21	0.037	0.013	0.004	0.023	0.002	0.007	0.815	0.153	0.224	0.151	0.060	0.219
22	1.713	0.020	0.086	0.005	0.001	0.005	-0.571	-0.242	0.074	0.033	0.044	0.163
23	-3.354	-17.324	4.792	4.081	34.998	128.592	-31.818	-0.234	32.941	2.144	0.343	1.259
24	2.063	0.046	0.167	0.013	0.005	0.019	1.730	0.001	0.093	0.006	0.000	0.000
25	-0.521	-0.038	0.043	0.013	0.004	0.016	-0.360	0.007	0.029	0.013	0.001	0.003
26	0.027	0.006	0.002	0.016	0.001	0.003	-0.724	-0.003	0.047	0.008	0.000	0.001
27	0.344	0.047	0.063	0.067	0.012	0.045	-0.987	-0.009	0.127	0.033	0.002	0.006
28	-0.440	-0.044	0.046	0.022	0.007	0.024	0.156	0.051	0.034	0.097	0.016	0.058
29	0.405	0.005	0.021	0.005	0.000	0.001	-0.285	-0.004	0.011	0.003	0.000	0.001

TABLE III SIX OUTLIER MEASURES BASED ON OLS AND MM ESTIMATES FOR DATA SET WITH ONE OUTLIER

	Least Square-estimate						MM-estimate					
ut of oints ex	t_i 3.000	d_i 3.000	CD_i 1.000	p_{ii} 0.243	$DFFITS_i$ 0.775	C_i 2.000	t_i 3.000	d_i 3.000	CD_i 1.000	p_{ii} 0.335	$DFFITS_i$ 0.775	C_i 2.000
1	3.810	0.741	0.643	0.085	0.216	0.515	9.515	0.000	1.155	0.044	1.78E-05	4.25E-05
2	-0.907	-0.138	0.166	0.100	0.044	0.104	-1.357	0.000	0.192	0.060	5.13E-05	1.22E-04
3	-0.386	-0.061	0.075	0.114	0.020	0.049	-0.062	0.000	0.010	0.079	1.11E-04	2.63E-04
4	0.110	0.019	0.022	0.125	0.007	0.016	1.170	-0.001	0.214	0.100	1.82E-04	4.33E-04
5	-0.406	-0.070	0.085	0.131	0.025	0.061	0.017	-0.001	0.003	0.120	3.64E-04	8.66E-04
6	-0.442	-0.074	0.093	0.132	0.027	0.064	-0.043	-0.002	0.009	0.137	5.90E-04	1.40E-03
7	-0.163	-0.025	0.033	0.127	0.009	0.021	0.608	-0.002	0.134	0.146	7.27E-04	1.73E-03
8	-0.394	-0.053	0.078	0.119	0.018	0.043	-0.010	-0.003	0.002	0.146	1.22E-03	2.89E-03

TABLE III(CONTINUE). SIX OUTLIER MEASURES BASED ON OLS AND MM ESTIMATES FOR DATA SET WITH ONE OUTLIER

9	-0.012	-0.001	0.002	0.112	0.000	0.001	0.787	-0.003	0.170	0.139	9.72E-04	2.31E-03
10	-0.695	-0.077	0.135	0.113	0.026	0.061	-1.015	-0.329	0.216	0.135	1.21E-01	2.88E-01
11	-0.731	-0.087	0.149	0.124	0.030	0.072	-1.310	-0.330	0.286	0.143	1.25E-01	2.97E-01
12	0.307	0.042	0.068	0.146	0.016	0.038	0.933	-0.002	0.219	0.166	6.29E-04	1.50E-03
13	0.877	0.148	0.211	0.173	0.062	0.146	2.064	0.004	0.529	0.197	1.64E-03	3.91E-03
14	0.659	0.132	0.167	0.193	0.058	0.138	1.342	0.004	0.362	0.218	2.05E-03	4.87E-03
15	0.112	0.023	0.029	0.195	0.010	0.024	-0.093	-0.033	0.025	0.211	1.52E-02	3.61E-02
16	-0.160	-0.032	0.039	0.179	0.013	0.032	-0.765	-0.174	0.190	0.185	7.47E-02	1.78E-01
17	0.088	0.015	0.021	0.169	0.006	0.015	-0.095	-0.030	0.023	0.172	1.25E-02	2.98E-02
18	-0.290	-0.059	0.076	0.208	0.027	0.064	-0.810	-0.123	0.219	0.219	5.77E-02	1.37E-01
19	-0.362	-0.133	0.130	0.386	0.082	0.196	-0.687	-0.075	0.248	0.392	4.70E-02	1.12E-01
20	0.260	0.252	0.152	1.028	0.255	0.608	1.225	-0.002	0.657	0.863	2.25E-03	5.35E-03

TABLE IV SIX OUTLIER MEASURES BASED ON OLS AND MM ESTIMATES FOR DATA SET WITH 3 OUTLIERS (CASE 6,7,8)

	Least Square-estimate						MM-estimate					
Cut of points Index	t_i	d_i	CD_i	p_{ii}	$DFFITS_i$	C_i	t_i	d_i	CD_i	p_{ii}	$DFFITS_i$	C_i
	3.000	3.000	1.000	0.235	0.775	2.000	3.000	3.000	1.000	0.335	0.775	2.000
1	-0.704	-0.101	0.167	0.170	0.042	0.099	-0.117	-9.65E-05	0.014	0.044	2.03E-05	4.83E-05
2	-1.141	-0.167	0.271	0.169	0.069	0.164	-1.079	-2.19E-04	0.153	0.060	5.35E-05	1.27E-04
3	-0.807	-0.107	0.188	0.163	0.043	0.103	-0.050	-4.32E-04	0.008	0.079	1.21E-04	2.89E-04
4	-0.494	-0.059	0.111	0.151	0.023	0.055	0.930	-6.17E-04	0.170	0.100	1.95E-04	4.64E-04
5	-0.880	-0.097	0.187	0.136	0.036	0.085	0.013	-1.15E-03	0.003	0.121	4.00E-04	9.51E-04
6	1.992	0.232	0.398	0.120	0.080	0.192	7.978	-1.53E-03	1.706	0.137	5.66E-04	1.35E-03
7	2.156	0.245	0.406	0.107	0.080	0.191	8.526	-2.17E-03	1.882	0.146	8.31E-04	1.98E-03
8	1.989	0.216	0.362	0.099	0.068	0.162	8.033	-2.97E-03	1.771	0.146	1.13E-03	2.70E-03
9	-0.626	-0.068	0.115	0.101	0.022	0.051	0.623	-3.04E-03	0.134	0.139	1.14E-03	2.71E-03
10	-1.069	-0.145	0.206	0.111	0.048	0.115	-0.809	-3.12E-01	0.172	0.135	1.15E-01	2.73E-01
11	-1.052	-0.168	0.218	0.129	0.060	0.143	-1.044	-3.59E-01	0.228	0.143	1.36E-01	3.23E-01
12	-0.274	-0.045	0.061	0.149	0.017	0.041	0.738	-2.84E-03	0.174	0.166	1.16E-03	2.75E-03
13	0.196	0.034	0.046	0.164	0.014	0.032	1.636	5.29E-03	0.420	0.197	2.35E-03	5.60E-03
14	0.137	0.023	0.032	0.168	0.009	0.022	1.062	2.19E-03	0.286	0.218	1.02E-03	2.44E-03
15	-0.134	-0.020	0.031	0.159	0.008	0.019	-0.079	-4.59E-02	0.021	0.211	2.11E-02	5.02E-02
16	-0.215	-0.029	0.047	0.146	0.011	0.026	-0.613	-1.92E-01	0.152	0.185	8.27E-02	1.97E-01
17	0.054	0.008	0.012	0.150	0.003	0.007	-0.081	-3.28E-02	0.019	0.172	1.36E-02	3.25E-02
18	-0.112	-0.023	0.029	0.202	0.010	0.024	-0.648	-1.50E-01	0.175	0.219	7.04E-02	1.68E-01
19	-0.068	-0.028	0.024	0.372	0.017	0.040	-0.550	-7.21E-02	0.199	0.392	4.51E-02	1.07E-01
20	0.468	0.439	0.255	0.887	0.414	0.986	0.969	-3.65E-03	0.519	0.862	3.39E-03	8.07E-03

VI. CONCLUSION

In this paper, a linear approximation of a nonlinear model is formulated and subsequently leverage matrix based on the gradient is formed. The outlier measures for nonlinear regression are then formulated by incorporating the leverage matrix and the commonly used detection measures based on the OLS, M and MM estimates. The results of the study clearly reveal that the proposed measures which are based on the OLS and M estimates can hardly detect the high leverage points correctly. The studentized residuals-OLS based and CD-M based can detect a single outlier correctly while the t_i-M based able to detect 3 outliers correctly. The results of simulation study agree reasonably well with the results of the real data that the ti-MM based and CD-MM based are the best outlier measures in nonlinear regression because they consistently can identify outliers correctly in different outliers scenarios.

TABLE V SIX OUTLIER MEASURES BASED ON THE OLS AND MM ESTIMATES WITH 6 HIGH LEVERAGE POINTS(THE LAST 6 OBSERVATIONS)

	Least Square-estimate						MM-estimate					
Cut of points	t_i	d_i	CD_i	p_{ii}	$DFFITS_i$	C_i	t_i	d_i	CD_i	p_{ii}	$DFFITS_i$	C_i 2.000
Index	3.000	3.000	1.000	0.241	0.679	2.000	3.000	3.000	1.000	0.298	0.679	
1	-0.841	-0.027	0.095	0.038	0.005	0.015	-0.111	-8.30E-05	0.010	0.022	1.23E-05	3.41E-05
2	-1.709	-0.070	0.209	0.045	0.015	0.041	-1.026	-1.98E-04	0.105	0.031	3.50E-05	9.68E-05
3	-0.858	-0.038	0.113	0.052	0.009	0.024	-0.047	-3.67E-04	0.006	0.043	7.65E-05	2.12E-04
4	-0.038	-0.002	0.005	0.060	0.000	0.001	0.879	-5.62E-04	0.123	0.059	1.36E-04	3.77E-04
5	-0.792	-0.045	0.119	0.067	0.012	0.032	0.011	-9.74E-04	0.002	0.077	2.70E-04	7.47E-04
6	-0.772	-0.049	0.122	0.075	0.013	0.037	-0.035	-1.48E-03	0.006	0.096	4.59E-04	1.27E-03
7	-0.232	-0.016	0.038	0.081	0.005	0.013	0.456	-1.89E-03	0.089	0.115	6.41E-04	1.77E-03
8	-0.482	-0.035	0.082	0.086	0.010	0.029	-0.012	-3.00E-03	0.003	0.129	1.08E-03	2.98E-03
9	0.265	0.020	0.046	0.090	0.006	0.017	0.595	-2.89E-03	0.127	0.136	1.07E-03	2.95E-03
10	-0.682	-0.054	0.119	0.091	0.016	0.045	-0.785	-2.58E-01	0.166	0.135	9.47E-02	2.62E-01
11	-0.588	-0.046	0.102	0.090	0.014	0.039	-1.006	-2.46E-01	0.208	0.128	8.80E-02	2.44E-01
12	1.193	0.097	0.203	0.087	0.029	0.079	0.690	-3.08E-03	0.139	0.122	1.08E-03	2.98E-03
13	2.169	0.191	0.363	0.084	0.055	0.153	1.519	3.76E-03	0.307	0.122	1.32E-03	3.65E-03
14	1.852	0.160	0.308	0.083	0.046	0.128	0.978	6.04E-04	0.205	0.132	2.19E-04	6.08E-04
15	0.955	0.084	0.163	0.088	0.025	0.069	-0.083	-3.01E-02	0.018	0.147	1.16E-02	3.20E-02
16	0.367	0.037	0.068	0.103	0.012	0.033	-0.594	-1.90E-01	0.138	0.163	7.67E-02	2.12E-01
17	0.454	0.060	0.096	0.135	0.022	0.061	-0.086	-1.96E-02	0.021	0.172	8.11E-03	2.25E-02
18	-0.632	-0.117	0.161	0.194	0.052	0.143	-0.620	-1.39E-01	0.148	0.171	5.76E-02	1.60E-01
19	-1.431	-0.406	0.448	0.294	0.220	0.609	-0.492	-5.99E-02	0.114	0.162	2.42E-02	6.69E-02
20	-1.497	-0.644	0.586	0.459	0.437	1.209	0.728	-3.01E-03	0.163	0.150	1.16E-03	3.22E-03
21	0.648	0.124	0.166	0.197	0.055	0.153	31.062	-1.64E-04	7.779	0.188	7.13E-05	1.98E-04
22	-0.056	-0.011	0.014	0.200	0.005	0.013	31.156	-1.35E-04	7.838	0.190	5.89E-05	1.63E-04
23	-1.206	-0.263	0.324	0.216	0.122	0.338	30.293	-1.22E-04	7.636	0.191	5.34E-05	1.48E-04
24	-0.349	-0.074	0.094	0.216	0.034	0.095	31.242	-1.22E-04	7.875	0.191	5.34E-05	1.48E-04
25	1.356	0.269	0.347	0.197	0.119	0.330	31.852	-1.64E-04	7.976	0.188	7.13E-05	1.98E-04
26	-0.127	-0.030	0.036	0.243	0.015	0.041	31.880	-1.11E-04	8.051	0.191	4.85E-05	1.34E-04

REFERENCES

[1] Anskombe, F. J. and Tukey, J. w. (1963), The examination and analysis of residuals. Technometrics, 5, 141-60.
[2] Atkinson, A.C., (1981), Two graphical displays for outlying and influential observations in regression, Biometrika, 68, 1, 13-20.
[3] Atkinson, A.C., (1982), Regression Diagnostics, Transformations and Constructed Variables, Journal od Royal Statistical Society, B, 44, 1, 1-36.
[4] Atkinson, A.C., (1986). Masking unmasked, Biometrika, 73, 3, 533-541.
[5] Bates, D.M. Watts, D.G., (1980). Relative curvature measures of nonlinearity, J. R. statist. Ser. B 42, 1-25.
[6] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), Regression Diagnostics, John Wiley & Sons, New York.
[7] Cook, R. D., and Weisberg, S., (1982), Residuals and Influence in Regression. CHAPMAN and HALL.
[8] Fox, T., Hinkley, D. and Larntz, K., (1980), Jackknifing in nonlinear regression. Technometrics, 22, 29-33.
[9] Habshah, M., Noraznan, M. R., Imon, A. H. M. R. (2009). The performance of diagnostic-robust generalized potential for the identification of multiple high leverage points in linear regression, Journal of Applied Statistics, 36(5):507-520.
[10] Hadi, A.H. (1992). A new measure of overall potential influence in linear regression, Computational Statistics and Data Analysis 14 (1992) 1-27.
[11] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), Robust Statistics: The Approach Based on InfluenceFunctions. New York: John Wiley & Sons, Inc.
[12] Hoaglin, D. C., Mosteller, F., Tukey, J. W. (1983), Understanding Robust and Exploratory Data Analysis, John Wiley and Sons.
[13] Hoaglin, D.C., & Wellsch, R. (1978). The hat Matrix in regression and ANOVA. Ammerican Statistician 32, 17-22.
[14] Huber, P. J. (1981), Robust Statistics, Wiley, New York .
[15] Imon, A.H.M.R, (2002), Identifying multiple high leverage points in linear regression, J. Stat. Stud. 3, 207–218.
[16] Kennedy. W. and Gentle, J. (1980). Statistical Computing. New York:Dekker.
[17] Rousseeuw, P. J., and Leroy, A. M. (1987), Robust Regression and outlier detection, New York: John Wiley.
[18] Riazoshams, H., Habshah, Midi, (2009), A Nonlinear regression model for chickens' growth data. European Journal of Scientific Research, 35, 3, 393-404.
[19] Srikantan, K. S. (1961), Testing for a single outlier in a regression model. Sankhya A, 23, 251-260.
[20] Stromberg, A. J., (1993), Computation of High Breakdown Nonlinear Regression Parameters, Journal of American Statistical Association, 88 (421), 237-244.
[21] Seber, G., A. F. and Wild, C. J. (2003), Nonlinear Regression, John Wiley and Sons.
[22] Yohai, V. J. (1987), High Breakdown point and high efficiency robust estimates for regression, The Annals of Statistics, 15, 642-656.