Jurnal Teknologi, 41(C) Dis. 2004: 11–26

THE RESIDUAL PLOT FOR A NON-LINEAR REGRESSION MODEL WITH THE PRESENCE OF OUTLIERS AND HETEROSCEDASTIC ERRORS

HABSHAH MIDI[1] & AZMI JAAFAR[2]

Abstract. Robust regression is extremely useful in assessing the adequacy of a fit and suggesting appropriate transformations. This can be achieved in a single run by using robust estimation methods instead of constructing outlier diagnostics. In this paper, the performance of the residual plot of the robust Weighted MM estimators (WMM) was compared with the Non-Linear Least Squares (NLLS), Generalized Non-Linear Least Squares (GNLLS), and robust MM residual plots. The results obtained from numerical examples signified that the residual plots from the NLLS and the GNLLS fit can hardly identify outliers and high leverage points. Furthermore, it did not show an ideal pattern in the residuals within [–2.5, 2.5] interval. The GNLLS and the WMM residual plots revealed an ideal pattern when the error variances were heteroscedastic and no contamination occured in the data set. The residual plot from the MM fit can identify outliers but the residuals within the [–2.5, 2.5] interval indicated that the residuals increased with increasing estimate response, which may suggest an appropriate transformation. The WMM residual plot exhibited a pronounced ideal pattern denoted by its residuals, which were randomly distributed within [–2.5, 2.5] interval.

Keywords: Outliers, robust regression, heteroscedastic, weighted MM estimator

Abstrak. Regresi teguh adalah sangat berguna bagi menilai kecukupan satu penyesuaian dan mencadangkan penjelmaan yang sesuai. Ini boleh dicapai dalam hanya satu pelaksanaan penganggaran teguh dan bukannya membina satu diagnostik titik terpencil. Dalam kertas ini, prestasi plot reja bagi Penganggar Teguh MM Berpemberat (WMM) dibandingkan dengan plot reja Kuasadua Terkecil Tak Linear (NLLS), Kuasadua Terkecil Tak Linear Teritlak (GNLLS) dan Penganggar Teguh MM. Dari keputusan berangka yang diperoleh, menyatakan bahawa plot reja NLLS dan GNLLS sukar mengenal pasti titik terpencil dan titik pelarasan tinggi. Lagipun, ianya tidak menunjukkan corak impian dalam selang reja [–2.5, 2.5]. Plot reja GNLLS dan WMM menghasilkan corak impian apabila varian ralat adalah heterosedastik dan tiada kekotoran berlaku dalam data set. Plot reja dari MM boleh mengenal pasti titik terpencil tetapi reja dalam selang [–2.5, 2.5] menyatakan bahawa ianya menokok dengan menokoknya tindakbalas penganggar, yang mungkin memerlukan penjelmaan yang sesuai. Plot reja WMM mempamerkan corak impian unggul dengan rejanya yang bertaburan secara rawak di dalam selang [–2.5, 2.5].

Kata kunci: Titik terpencil, regresi teguh, heterosedastik, penganggar MM berpemberat

1 Laboratory of Statistics and Applied Mathematics, Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400 UPM Serdang.
2 Department of Mathematics, Universiti Putra Malaysia, 43400 UPM Serdang.

1.0 INTRODUCTION

Outlier is a single or a group of observations which are markedly different from the bulk of the data or from the pattern set by the majority of the observations. There are two types of outliers; i.e. outliers in the response variables and outliers in the explanatory variables. A leverage point is a point in the explanatory variable that lies far away from the bulk of the observed x_i's in the sample. The Ordinary Least Squares (OLS) method is often used in practice to estimate the parameters of a model. This procedure is less efficient when the errors come from some heavy tailed distribution, which may be caused by the presence of outliers. To remedy this problem, a robust (resistant) method is put forward. A robust regression is extremely useful in identifying outliers and assessing the adequacy of a fit and suggesting suitable transformations. All of these aspects can be detected in a single run by simply running a robust estimator. There are considerable papers related to robust regression such as [1-4]. An alternative approach in dealing with outliers in regression analysis is to construct outliers diagnostics. These are quantities computed from the data with the purposed of pinpointing influential observations, which can then be studied and corrected or deleted, and followed by a NLLS analysis on the 'good' data. Diagnostic and robust regression has the same goal, but they proceed in the opposite order. In a diagnostic setting, one first wants to identify the outliers and then, fit the good data in the classical way, whereas the robust approach first fits a regression that does justice to the majority of the data and then, discovers the outliers as those points having large residual plots from the robust fit. Midi [5] has displayed that the WMM estimator provides the most robust and efficient estimates compared to the NLLS, GNLLS, and MM estimator of a non-linear model in the presence of outliers and heteroscedastic errors. When the errors have a constant variance, we called this situation as homoscedasticity in contrast to heteroscedasticity in which, the variance of the errors are not constant. In this paper, we want to show that residual plots from the existing methods, such as the NLLS, GNLLS, and MM have lack of fit for non-linear model with heteroscedastic errors and would suggest for a suitable transformation. In addition, we also want to demonstrate that the residual plots associated with the WMM estimates give an ideal pattern, which is a horizontal band of points about the zero line.

2.0 A WMM ROBUST ESTIMATOR IN NON-LINEAR REGERSSION

Midi [8] has discussed that the WMM technique is computed in four stages:

(i) Compute the Weighted Non-linear Least Median Squares (LMS) [4-6].
(ii) Calculate an M estimate of weighted scale using rho function ρ_0.
(iii) Compute the weighted M estimate using rho function ρ_1.
(iv) Repeat steps 2 and 3 until convergence.

Hampel's redescending psi function [7], denoted as ρ_H is used in the analysis. Yohai [8] 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 [9] demonstrated that selecting $k_0 = 0.212$ and $k_1 = .9014$ will guarantee a high breakdown estimate and result in 95% efficiency under normal errors, respectively.

3.0 THE WEIGHTED NON-LINEAR LEAST MEDIAN OF SQUARES (LMS)

Consider the general heteroscedastic non-linear regression [10]:

$$y_i = f(x_i, \beta) + \alpha^2 g\{f(x_i, \beta), Z_i, \theta\} e_i \tag{1}$$

and the residuals,

$$r_i = \frac{y_i - f(x_i, \beta)}{g\{f(x_i, \beta), Z_i, \theta\}} \tag{2}$$

where β is the parameter model. The covariance matrix for y is the NXN matrix $\sigma^2 g\{f(x_i, \beta), Z_i, \theta\}$ to emphasize the possible dependence of the covariance matrix on the mean vector $f(x_i, \beta)$, the structural variance parameter θ and a matrix of known variables $Z = (z_1, z_2, \ldots z_n)$, whose individual component vectors z_i mayor may not include some or all of the predictors x_i. By assuming that variances are proportional to the regressor, the residual r_i in (2) becomes:

$$r_i = \frac{y_i - f(x_i, \beta)}{x_i} \tag{3}$$

For model (1), $\hat{\beta}_{WLMS}$ (Weighted Least Median Squares) is obtained from:

$$\underset{\beta_{WLMS}}{\arg\min}\, r^2_{(k)}\,(\beta_{WLMS}) \tag{4}$$

where

$$r_{(i)}{}^2\,(\beta_{WLMS}),\ i = 1,2,3,\ldots n \text{ are the ordered } r_i^2\,(\beta_{WLMS})$$

and $k = [[n/2\,]] + 1$, $[[.]]$ is the greatest integer function. The proposed algorithm for the Weighted Non-linear LMS is similar to Stromberg [9], except that y_i and $f(x_i, \beta)$ is replaced by y_i/x_i and $f(x_i, \beta)/x_i$, respectively.
The steps in the algorithm are as follows:

1. Calculate the initial estimate of WLMS denoted by $\hat{\beta}_{WLMS}$, using GNLLS denoted by $\hat{\beta}$ [10].
2. Compute the GNLLS estimate of p randomly selected points, denoted by $\hat{\beta}_{WLS}$ (Weighted Least Squares).
3. If the median squared residual at $\hat{\beta}_{WLS}$ is less than the median squared residual at $\hat{\beta}$, $\hat{\beta}$ is replaced by $\hat{\beta}_{WLS}$ as the current estimate of ($\hat{\beta}_{WLMS}$).
4. Steps 2 and 3 are repeated k times, where k is specified by Stromberg [9] and [11].
5. $\hat{\beta}$ is used as a starting value for calculating the LS (Least Squares) fit by $\hat{\beta}_{LS}^{*}$, for data points x_i such that $r_i^2(\hat{\beta}) \le \text{med}_{1\le i\le n}\, r_i^2(\hat{\beta})$. If $\text{med}_{1\le i\le n}\, r_i^2(\hat{\beta}_{LS}^{*}) < \text{med}_{1\le i\le n}\, r_i^2(\hat{\beta})$, then $\hat{\beta}$ is replaced by $\hat{\beta}^{*}_{LS}$ as the current estimate of $\hat{\beta}_{WLMS}$.
6. In order to get a still better estimate, the NeIder-Mead Simplex Algorithm [12] which is implemented in [13] with fractional tolerance 10^{-4}, is used to minimize $\text{med}_{1\le i\le n}\, r_i^2(\beta)$ by using $\hat{\beta}$ as the starting value.

4.0 THE WEIGHTED M-ESTIMATE FOR SCALE

Let $\hat{\beta}_{WLMS}$ be the parameter estimate of the regression function in (1) with a high breakdown point, and the residuals are defined by:

$$r_i(\hat{\beta}) = \frac{y_i - f(x_i, \hat{\beta}_{WLMS})}{x_i}, \qquad 1 \le i \le n$$

The weighted M-scale estimate is defined as the value of s which is the solution of

$$\frac{1}{n}\sum \rho_0(r_i / s_n) = b \tag{5}$$

where b may be obtained from the equation $E_\Phi(\rho(r)) = b$.
Let ρ_0 in (5) be a real function which satisfies the following assumptions;

(i) $\rho(0) = 0$
(ii) $\rho(-r) = \rho(r)$
(iii) $0 \le u \le v$ implies $\rho(u) \le \rho(v)$ (6)

(iv) ρ is continuous

(v) Let $a = \sup \rho(r)$, then $0 < a < \infty$

(vi) If $\rho(u) < a$ and $0 \leq u < v$, then $\rho(u) < \rho(v)$

The constant b is such that:

$$b/a = 0.5 \text{ where } a = \max \rho_0(r)a \tag{7}$$

This implies that this scale estimate has a breakdown point equals to 0.5 as verified by Huber [14].

5.0 WEIGHTED MM ESTIMATE

The weighted MM estimate [3,8] is found by minimizing:

$$S(\beta_{WMM}) = \Sigma\rho_1(r_i(\beta_{WLMS})/s_n) \tag{8}$$

where β_{WLMS} and s_n are defined in (4) and (5) respectively. Stromberg [11] interpreted $\rho_1(0/0)$ as 0. ρ_1 is another function which satisfies assumption (6) such that,

$$\rho_1(r) \leq \rho_0(r) \tag{9}$$

$$\sup \rho_1(r) = \sup \rho_0(r) = a \tag{10}$$

This implies that $S(\beta_{1WMM}) \leq S(\beta_{0WMM})$ where β_{1WMM} is the Weighted MM which will yield the specified efficiency under normal errors and β_{0WMM} are the Weighted MM with a high breakdown estimate.

6.0 A NUMERICAL EXAMPLE AND RESMUAL PLOTS

A numerical example are presented to examine the residual plots of the NLLS, GNLLS, MM, and the WMM estimates in the presence of outliers and leverage points. In order to assess these residual plots, 30 'good' data were generated according to Michaelis-Menten model [15].

$$y_i = \frac{\beta_1 x_i}{\exp(\beta_2) + x_i} + \varepsilon_i \tag{11}$$

where x_i are uniformly distributed on [0,10]. The errors ε_i were generated from a normal distribution N $(0, \sigma^2 x_i^2)$, $\sigma^2 = (.25)^2$, $\beta_1 = 10$, and $\beta_2 = 0$.

We studied four (4) data sets. The first set is the data set 1, which has been generated according to the above criteria. In the data set 2, we constructed 3 outliers in the y-direction by changing three y values in data set 1; i.e, $y_2 = 14.3$, $y_{12} = 16.4$, $y_{27} = 18.5$. We introduced outliers in the x direction in the data set 3. Now, we deleted 3 values of

x and y for observations 2,12,27 in data set 2 and replaced with $x_2 = 16.7$, $x_{12} = 20.3$, $x_{27} = 20.0$, $y_2 = 84.3$, $y_{12} = 86.4$, $y_{27} = 82.5$. Finally, in the data set 4, we constructed 3 high leverage outliers. All observations were as in the data set 3 except that we replaced y_2, y_{12}, y_{27} with values 254.3, 256.4, and 258.5 respectively.

The scatter plots of y versus x for the four data sets are shown in Figures 1-4. Figure 1 suggests that the values of x increases with the increasing values of y. This implies that there exists inequality of error variances; i.e. $\sigma^2(\varepsilon) = \sigma^2 x^2$. Outliers in the y direction, outliers in the x direction, and the high leverage outliers are illustrated in Figures 2,3 and 4, respectively. Table 1 provides the estimates of β along with the associated estimated variances for the four estimators discussed earlier. OT, LEV, and LOT, in

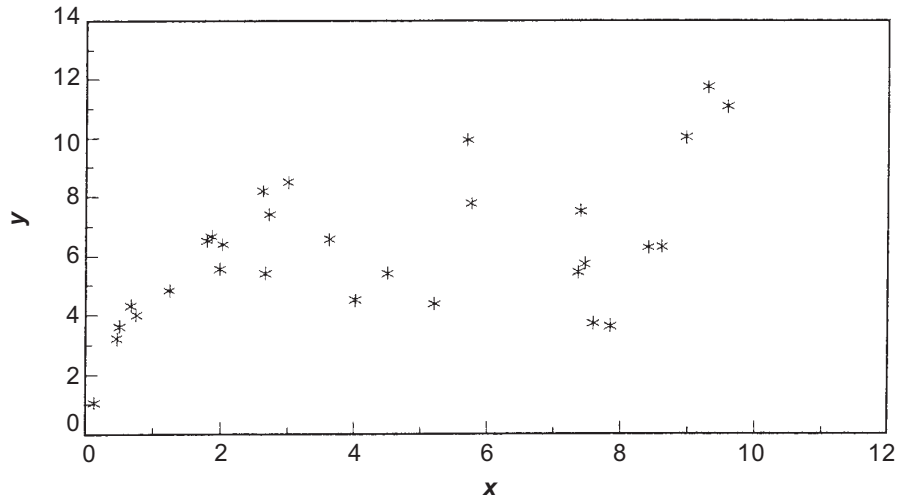

Figure 1 The scatter plots of y versus x

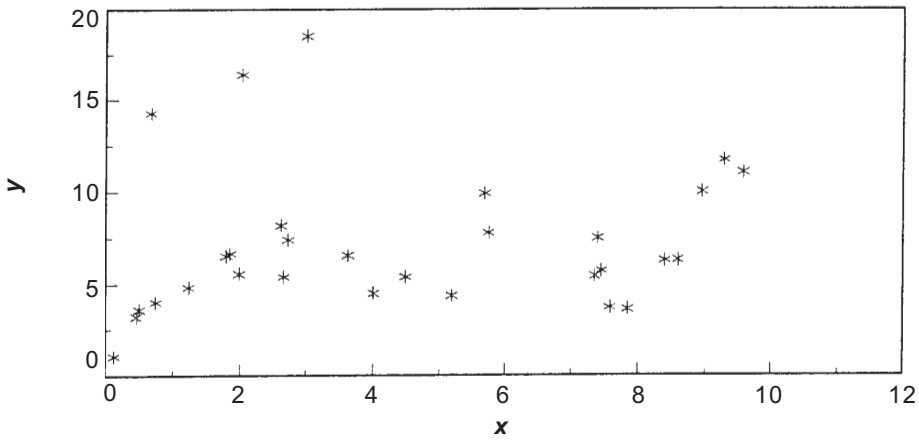

Figure 2 The scatter plots of y versus x (3 outliers)

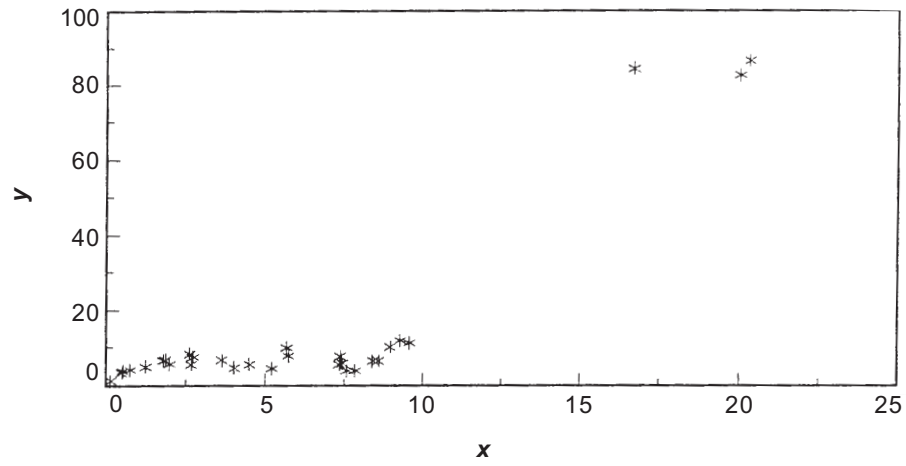

Figure 3 The scatter plots of y versus x (3 leverage points)

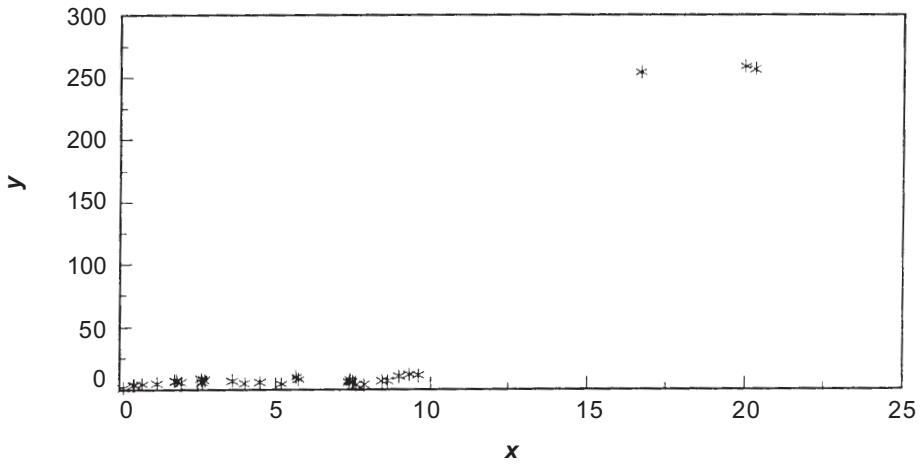

Figure 4 The scatter plots of y versus x (3 high leverage outliers)

Table 1 indicates outliers, leverage points, and high leverage points, respectively. The asterisks noted in the tables indicate that the estimated values exceed 1,000.

The results of Table 1 suggest that when there is no contamination in the data set, the GNLLS estimates give the best result. This is indicated by its smallest standard error among the four estimates. The WMM estimator emerges to be the best when compared to the other estimators in the situation where there are outliers and leverage points in the data set. The NLLS and the GNLLS estimates are seriously affected by the outliers. The MM estimates are slightly inferior to the WMM estimates. It is interesting to note that the GNLLS estimates are not influenced very much by the presence of leverage points. The explanation for this is that these extreme points may

Table 1 The value of $\hat{\beta}_1$ and $\hat{\beta}_2$ associated with their S.E.

Lev/outlier (#)	Method	$\hat{\beta}_1$	S.E ($\hat{\beta}_1$)	β_2	S.E ($\hat{\beta}_2$)
	NLLS	7.779	0.718	−0.540	0.524
	GNLLS	8.834	0.507	−181	0.094
0	MM	7.490	0.746	−0.676	0.603
	WMM	8.838	0.580	−0.181	0.107
	NLLS	8.006	1.041	−1.587	1.224
3 OT	GNLLS	13.030	4.172	0.019	0.514
	MM	7.399	0.881	−0.183	0.768
	WMM	8.547	0.573	−0.211	0.111
	NLLS	*****	*****	21.568	******
3 LEV	GNLLS	9.562	1.945	−0.049	0.321
	MM	7.455	.791	−0.603	0.702
	WMM	8.554	0.617	−0.209	0.116
	NLLS	*****	*****	20.568	*****
	GNLLS	960.204	11436.876	5.593	12.169
3 LOT	MM	7.455	0.791	−0.603	0.702
	WMM	8.547	0.601	−0.211	0.113

be considered as 'mild' leverage points, where the outlying *x* observations were quite close to the majority of the data.

Let us now examine the residual plots corresponding to the NLLS, GNLLS, MM, and WMM estimates for the four data sets. They are presented in Figures 5-19. According to Rousseeuw and Leroy [4], if the residual are normally distributed then one can expect that roughly 98% of the standardized residuals will lie in the interval [-2.5, 2.5). Thus, observation for which the standardized residuals are situated far from the horizontal band can be identified as outliers. Figures 5-8 show the residual plots when there was no contamination in the data. The plots in Figures 5 and 7 indicate that the residuals increased with increasing estimated response which indicated that the variance of the error terms were not constant. These plots suggest that the use of NLLS and MM estimators were not appropriate for the data sets and may turn to other suitable estimator. However, Figures 6 and 8 suggest an ideal pattern which signifies an adequate model and well behaved data. As can be expected, the GNLLS and WMM estimators are suitable for the heteroscedastic model.

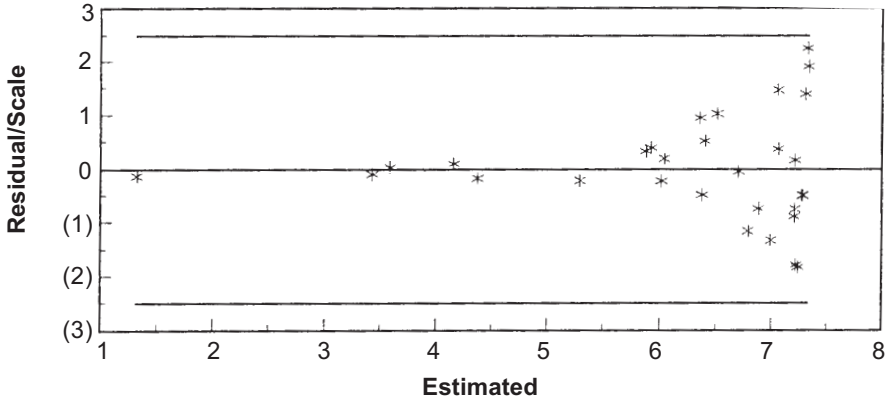

Figure 5 Std. residuals versus estimated values asso. with NLLS (No outliers)

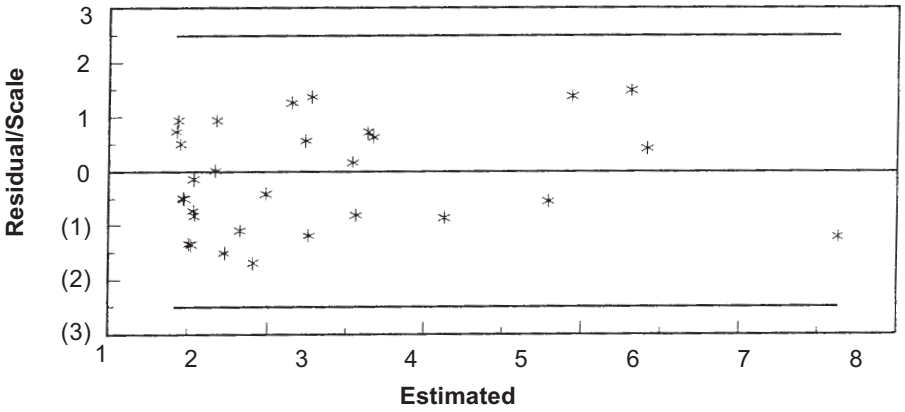

Figure 6 Std. residuals versus estimated values asso. with GNLLS (No outliers)

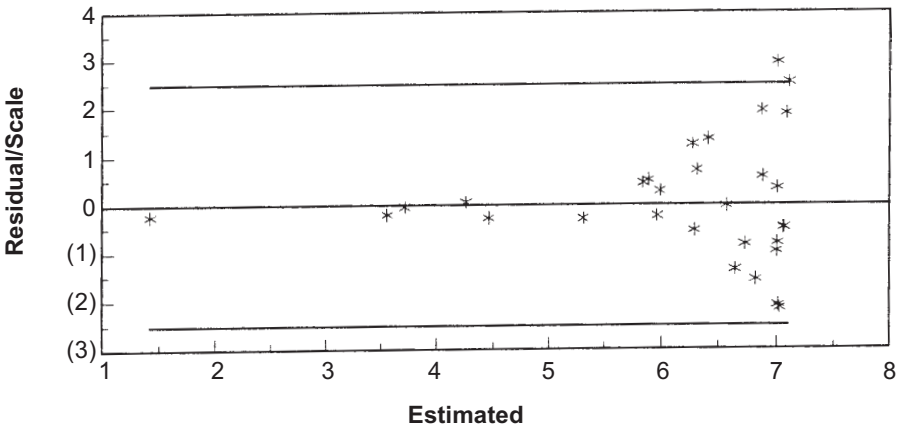

Figure 7 Std. residuals versus estimated values asso. with MM (No outliers)

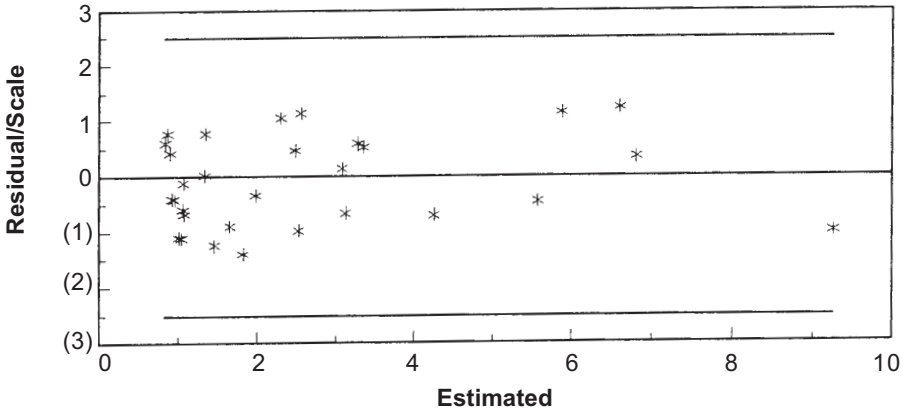

Figure 8 Std. residuals versus estimated values asso. with WMM (No outliers)

Figures 9-12 exhibit the residual plots when three outliers are present in the data set. In Figure 9, almost no outliers can be observed in the data. There are still anomalies in the pattern of the residuals associated with the NLLS estimators. The MM residual plot in Figure 11 identifies three outliers but does not show an adequate fit. The GNLLS plot of Figure 10 identified only one outlier and it does not show an ideal pattern in the residuals within [–2.5, 2.5) interval. It becomes immediately clear from Figure 12 associated with the VVMM residual plot that there are three outliers in the data set. This plot also reveals an ideal pattern which indicates an adequate model with residuals randomly distributed within the interval [–2.5, 2.5).

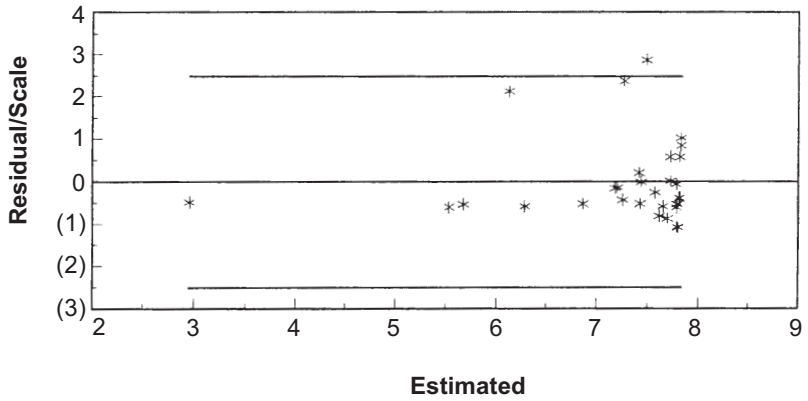

Figure 9 Std. residuals versus estimated values asso. with NLLS (3 outliers)

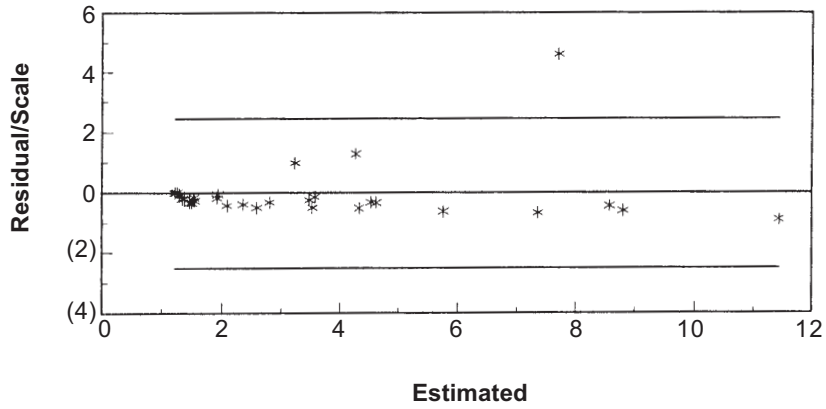

Figure 10 Std. residuals versus estimated values asso. with GNLLS (3 outliers)

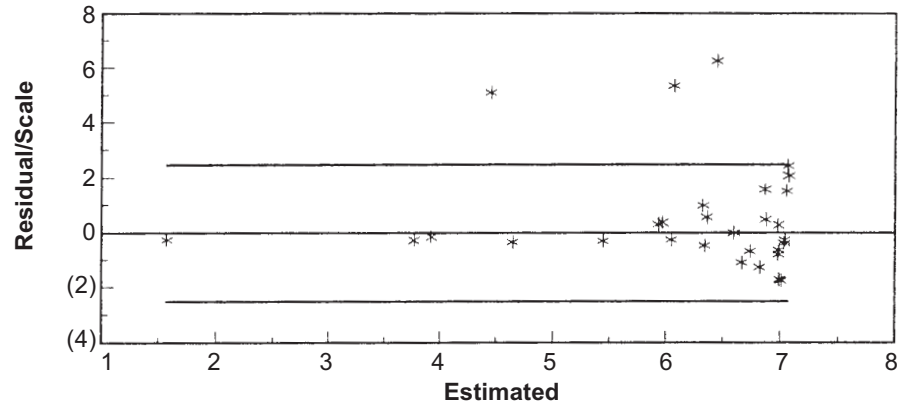

Figure 11 Std. residuals versus estimated values asso. with MM (3 outliers)

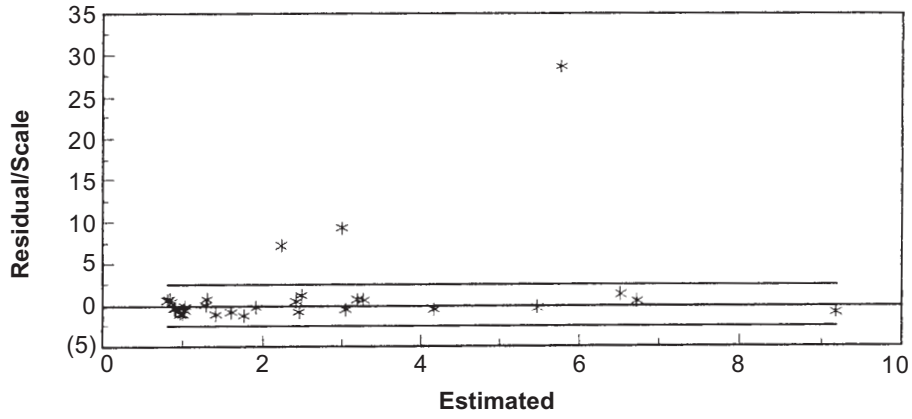

Figure 12 Std. residuals versus estimated values asso. with WMM (3 outliers)

The presence of leverage points changed the pattern of the residual plots of NLLS, as shown in Figure 13. It appeared as though there was no outlier. One has to keep in mind that leverage points tend to produce small NLLS residuals simply by virtue of their leverage. On the other hands, Figure 16, which is associated with the WMM plot gives evidence of the presence of 3 outlying observations and the residuals behave nicely within the interval [–2.5, 2.5). It is interesting to note that the residual plot in Figure 14 associated with the GNLLS with 3 leverage points shows an ideal pattern. However, the 3 leverage points identified by the GNLLS are intermediate because it is on the verge of the area containing the outliers. The MM residual plot in Figure 15 may give a misleading conclusion since the residuals appear not to be randomly distributed around the zero line. Figure 17-19 present the residual plots associated with three high leverage points. The presence of the high leverage points complicates the situation even more. From the residual plots in Figures 17 and 18, it is evident that the estimates of NLLS and GNLS are imperfect because the plots reveal an unexpected pattern of the residuals within the interval [–2.5, 2.5]. In order to be sure that this pattern is not caused by the presence of outliers, we will compare these figures to the residuals associated with the MM fit (Figure 15). The residuals inside the interval [–2.5, 2.5], tend to follow a fan shape which hints that transformation is necessary. The presence of high leverage points does not alter the pattern of the residual plots of the WMM estimates. It depicted an ideal pattern and identified three high leverage points (Figure 19).

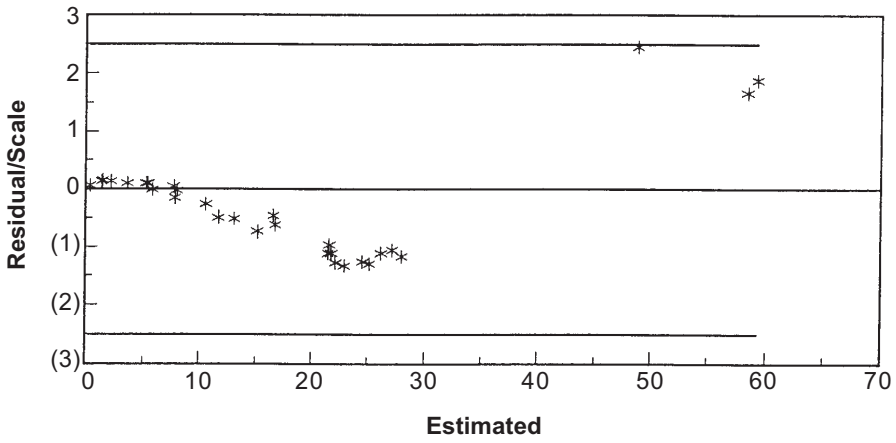

Figure 13 Std. residuals versus estimated values asso. with NLLS (3 leverage points)

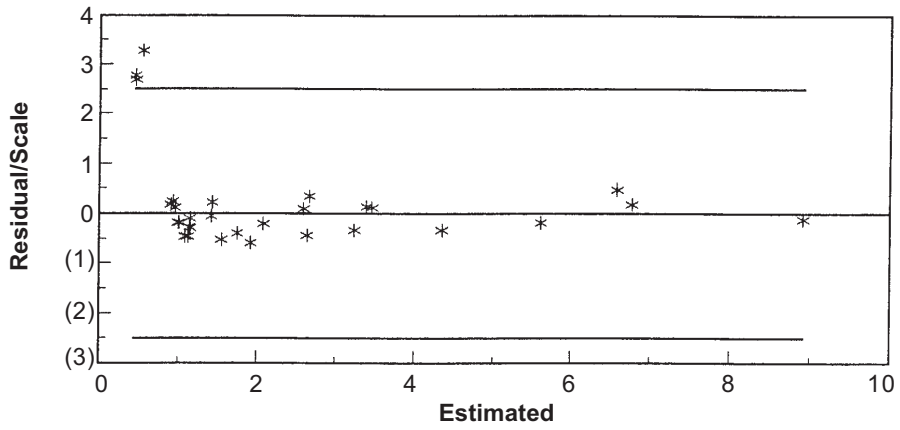

Figure 14 Std. residuals versus estimated values asso. with GNNLS (3 leverage points)

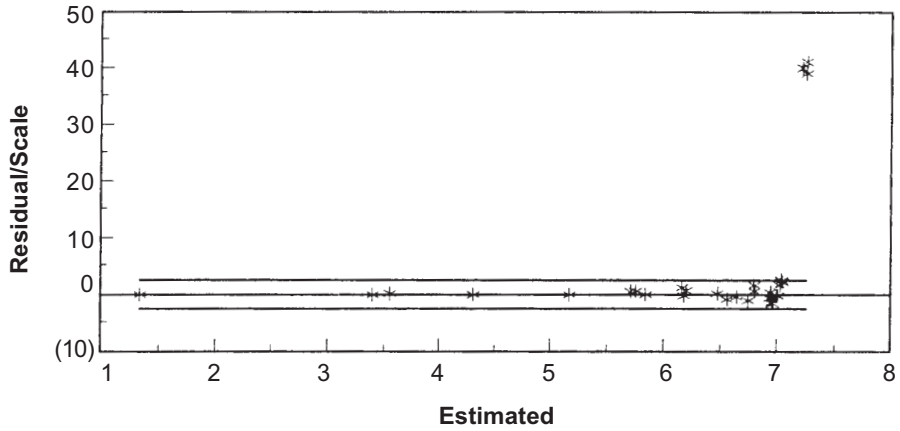

Figure 15 Std. residuals versus estimated values asso. with MM (3 leverage points)

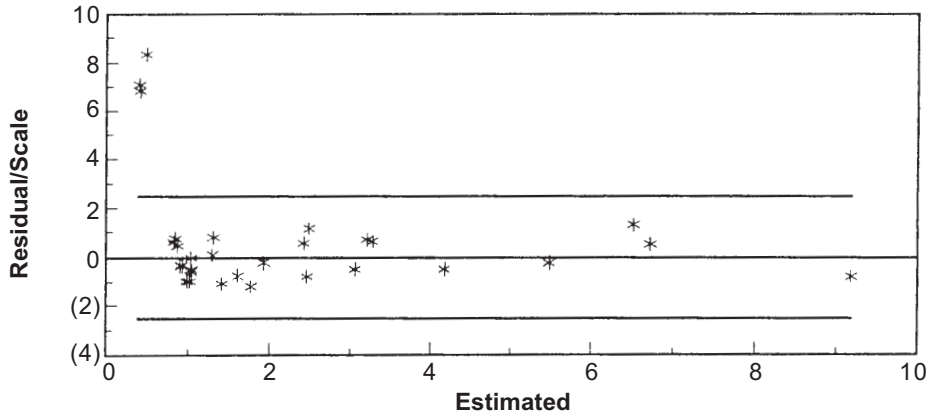

Figure 16 Std. residuals versus estimated values asso. with WMM (3 leverage points)

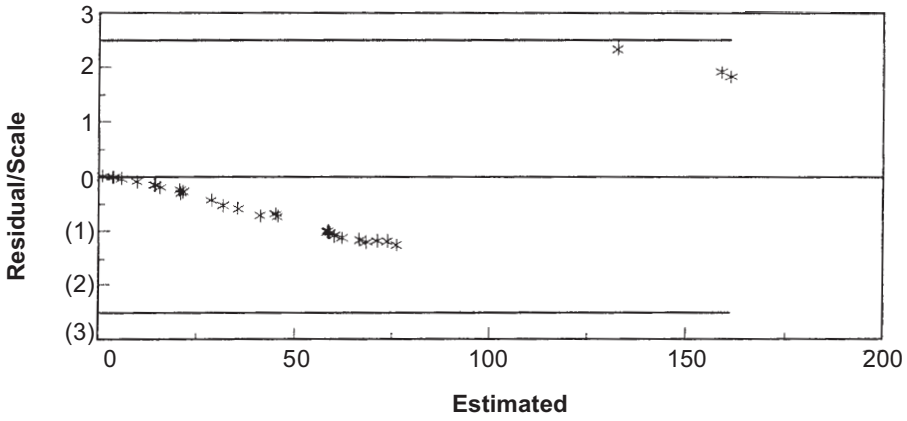

Figure 17 Std. residuals versus estimated values asso. with NLLS (3 high lev. outliers)

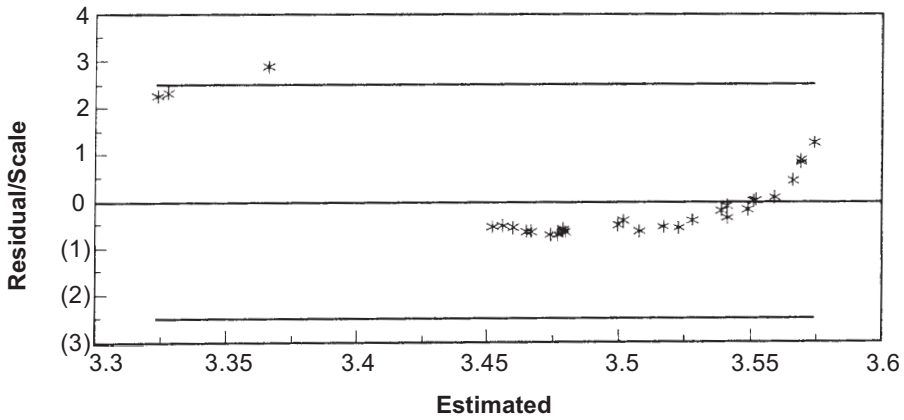

Figure 18 Std. residuals versus estimated values asso. with GNNLS (3 high lev. ourliers)

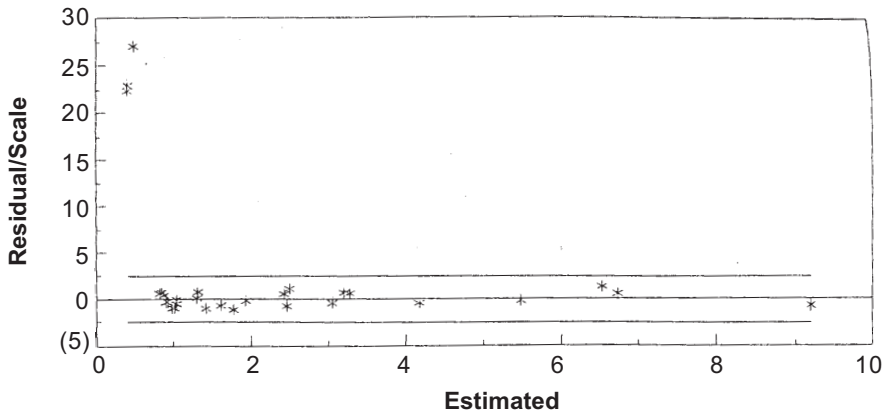

Figure 19 Std. residuals versus estimated values asso. with WMM (3 high lev. outliers)

7.0 CONCLUSIONS

From the numerical examples, it signified that the residual plots from the NLLS and the GNLLS fit can hardly identify outliers and high leverage points. Furthermore, it did not show an ideal pattern in the residuals within interval [–2.5, 2.5]. An ideal pattern of the residual plot which indicates an adequate model and a well-behaved data is when the points are scattered randomly around 0 without showing any obvious trend or shape. Anomalies in the pattern of the residuals suggested that the model was not appropriate and several courses of action should be taken. The GNLLS and the WMM residual plots revealed an ideal pattern when the error variances were heteroscedastic and no contamination occured in the data set. The residual plot from the MM fit can identify outliers but the residuals within the interval [–2.5, 2.5] indicated that the residuals increase with increasing estimated response, which may suggest an appropriate transformation. The WMM residual plot exhibited a pronounced ideal pattern denoted by its residuals, which were randomly distributed around 0 within residual interval [–2.5, 2.5]. The numerical examples clearly showed that the WMM was the best among all the other estimators because beside being able to detect the true outliers, there was no anomalies in the pattern of the residuals which implied that the WMM was the most suitable estimator for the data. Furthermore, in the presence of outliers, the standard errors of the WMM were the smallest compared to the NLLS, GNLLS, and MM estimators.

REFERENCES

[1] Armstrong, R. D., and M. T. Kung. 1978. Least Absolute Value Estimates for a Simple Linear Regression Problems. *Applied Statistics.* 27: 363-366.

[2] Barrodale, I., and F. D. K. Roberts. 1973. An Improved Algorithm for Discrete L-1 Linear Approximation. *Siam Journal of Numerical Analysis*. 10: 839-848.

[3] Carrol, R. J., and D. Ruppert. 1982. Robust Estimation in Heteroscedastic Linear Models. *Ann. Stat*. 10: 429-41.

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

[5] Midi, H. 1998. Robust Non-linear Regression Estimator with Heteroscedastic Errors. *Int. Sc.* 11 (4)

[6] Rousseeuw, P. J. 1984. Least Median of Squares Regression. *J. Am. Stat. Assoc*. 79: 871-879.

[7] Hampel, F. R., E. M. Ronchetti., P. J. Rousseeuw., and W. A. Stahel. 1986. *Robust Statistics: The Approach Based On Influence Functions*. New York: Wiley.

[8] Yohai, V. J. 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *Ann. Statist*. 15(20): 642-656.

[9] Stromberg, A. J. 1993. Computational of High Breakdown Non-linear Regression Parameters. *J. Am. Stat. Assoc.* 88 (421): 237-244.

[10] Carrol, R. J., and D. Ruppert. 1988. *Transformations and Weighting in Regression*. London: Chapman and Hall.

[11] Stromberg, A. J. 1992. High Breakdown Estimators in Non-linear Regression. *In L_1-Statistical Analysis and Related Methods*. Ed. Y. Dodge. Amsterdam: North-Holland. 103-112.

[12] Nelder, J., and R. Mead., 1965. A Simplex Method for Function Minimization. *Computer J.* 7: 308- 313.

[13] Press, W. H., B. P. Flannery., S. A. Teukolsky., and W. T. Vetterling. 1986. *Numerical Recipes: The Art of Scientific computing*. New York: Cambridge University Press.

[14] Huber, P. J.1981. *Robust Statistics.* New York: Wiley.

[15] Ratkowsky, D. A. 1983. Non-linear Regression Modelling. New York: Marcel Dekker.