مجلة العلوم البحثة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 03/02/2019 Revised 15/04/2019 Published online 01/05/2019

A comparative study of some robust nonlinear regression methods

*Abdelgadir Khalifa Alsalem[a], Alsaidi M. Altaher[b]
[a]Department of Mathematical sciences, Basic Science School/Libyan Academy, Libya
[b]Department of Statistics, Faculty of Science/University of Sebha, Libya
*Corresponding author: abdelgadirkhalifa@yahoo.com

Abstract Least squares (LS) with Gauss-Newton method is the most widely used approach to estimate the parameters of nonlinear regression models. In the presence of outliers, even one single unusual value may have a large effect on the parameter estimates. This paper aims to introduce some popular robust nonlinear techniques that commonly used as a better alternative method to the classical least squares. This includes M-estimator and MM. In addition, the target is to compare their practical performance under a variety of circumstances such as sample size, percentage of outliers and model formula. Results of Monte Carlo simulations and real data example using R software, indicated that the best performance has been achieved by MM followed by M estimator for all possible percentages of outliers (10%, 20%, 30%, 40%) as well as all sample sizes (n=50, n=100, and n=150). Moreover, results approved that the LS estimator remains the best when there is no outlier in data.
Keywords: robust estimator, nonlinear model, outliers, M-estimator, MM-estimator.

مقارنة بين بعض الطرق المتينة في الانحدار اللا خطي

*عبدالقادر خليفة السالم[1] و السعيدي المهدي الطاهر[2]
[1] قسم العلوم الرياضية– مدرسة العلوم الأساسية– الأكاديمية الليبية، ليبيا
[2] قسم الإحصاء– كلية العلوم–جامعة سبها، ليبيا
*للمراسلة:abdelgadirkhalifa@yahoo.com

الملخص طريقة المربعات الصغرى (LS) مع طريقة Gauss-Newton هي الطريقة الأكثر استخداماً لتقدير معلمات نماذج الانحدار الغير خطية. في وجود القيم المتطرفة، قد يكون حتى لمفردة واحدة متطرفة تأثيرا كبيرا على تقديرات المعالم. تهدف هذه الورقة إلى عرض بعض التقنيات غير الخطية المتينة الشائعة والتي تستخدم عادةً كطرق بديلة أفضل للمربعات الصغرى الكلاسيكية، وهذا يشمل مقدر M، ومقدر MM. بالإضافة إلى ذلك، فإن الهدف هو مقارنة أدائها العملي تحت ظروف مختلفة مثل حجم العينة ونسبة القيم المتطرفة وصيغة النموذج. أشارت نتائج المحاكاة والبيانات الحقيقية باستخدام برنامج R إلى أن أفضل أداء تم تحقيقه بواسطة MM متبوعا بتقدير M لجميع النسب المئوية المحتملة من القيم المتطرفة (10 ٪، 20 ٪، 30 ٪، 40 ٪) وكذلك جميع أحجام العينات (50، 100، 150). علاوة على ذلك، وافقت النتائج على أن مقدر LS يبقى الأفضل عند عدم وجود قيم متطرفة.
الكلمات المفتاحية: المقدر المتين، النموذج اللاخطي، القيم المتطرفة، مقدر M ،مقدر MM.

1. Introduction

Nonlinear regression is one of the most popular and widely used models in analyzing the effect of explanatory variables on a response variable when the underling regression function is nonlinear. It has many applications in scientific research such as in dose response studies conducted in agricultural sciences, toxicology and other biological sciences; see [1] and [2]. With the presence of outliers in the data, the ordinary least squares LS method provides misleading values for the parameters of the nonlinear regression, and predictions may no longer be reliable, see [3]. Outliers are those observations that deviate markedly from other members of the observations or data points which are unusually large or small from the majority of the observations. They are also called the abnormal data. Outliers can arise due to measurement or recording error, natural variation of the underlying distribution, or a sudden alteration in the operating system. One disadvantages of nonlinear least square with Gauss-Newton method method is its sensitivity to the presence of even few outlying observations. As a result, the errors in the process of prediction and estimating as they amplify the variance of errors, leading to extended the length of confidence interval and reduced estimation efficiency. The proposed solutions for estimating the parameters of the nonlinear regression model in the presence of outliers is the use of robust estimators rather than the method of nonlinear least squares. In statistical literature, most of the robust linear regression techniques are successfully adopted for nonlinear setting, such as M-estimator, MM-estimator, Least Median of Squares (LMS), Least Trimmed Squares (LTS), See [4], [5], [7], [8] and [9]. So much comparisons had been conducted to compare the robust estimators in linear regression, see for example [10], [11], and [12]. However, Little work has been done in nonlinear regression. Hence, such simulation comparisons in nonlinear regression are required. This work shall contribute in presenting new simulation results to determine which robust

estimator should be used when outliers are existing in data. To the best of the authors knowledge there is no practical comparison between M- estimator and MM- estimator.

The remainder of this paper is organized as follows: Section 2 gives a brief review to the Gauss- Newton Method, M-estimator, and MM-estimator. In section 3, simulation study and real data applications are present. Conclusions are drawn in section 4.

2. Methods and Models

This section is devoted to introduce brief theoretical descriptions of Gauss-Newton method for classical LS, M-estimator and MM-estimator.

2.1 The Gauss - Newton Method

A nonlinear regression model can be written as

$$y_i = f(x_i, \theta) + \varepsilon_i \quad , i = 1, \dots, n \qquad (1)$$

Where y is the dependent variable, f is a nonlinear function of the parameter $\theta.$, x_i is a vector of associated regressor variables for the *ith* observation . The error term, ε is assumed to be independent with constant variance, and usually that $\varepsilon \sim N(0, \sigma^2)$. To estimate the parameters of (1) using the Gauss – Newton method, a Taylor series expansion is used to approximate the nonlinear regression model with linear terms and then employs ordinary least squares to estimate the unknown parameters. The Gauss – Newton method begins with initial values for the regression parameters $\theta_0, \theta_1, \dots\dots, \theta_p$ denoted by $\theta^{(0)}$, where the superscript in parentheses denotes the iteration number. Once the starting values for the parameters have been chosen, the mean responses $f(X;\theta)$ is approximated for the n cases by the linear terms in the Taylor series expansion around the starting values $\theta_k^{(0)}$, such as:

$$f(x_i, \theta) \approx f\left(x_i, \theta^{(0)}\right) + \sum_{k=0}^{p} \left[\frac{\partial f(x_i, \theta)}{\partial \theta_k}\right]_{\theta=\theta^{(0)}} \left(\theta_k - \theta_k^{(0)}\right) \qquad (2)$$

where $\theta = (\theta_0, \theta_1, \dots\dots, \theta_p)', \theta^{(0)} = (\theta_0^{(0)}, \theta_1^{(0)}, \dots\dots, \theta_p^{(0)})'$

Note that $\theta^{(0)}$ is the vector of the parameter starting values.

Let

$$f_i^{(0)} = f(x_i, \theta^{(0)})$$
$$\beta_k^{(0)} = \theta_k - \theta_k^{(0)}$$
$$F_{ik}^{(0)} = \left[\frac{\partial f(x_i, \theta)}{\partial \theta_k}\right]_{\theta=\theta^{(0)}}$$

The Taylor approximation (2) becomes in this notation:

$$f(x_i, \theta) \approx f_i^{(0)} + \sum_{k=0}^{p} F_{ik}^{(0)} \beta_k^{(0)}$$

And an approximation to the nonlinear regression model (1) is:

$$Y_i \approx f_i^{(0)} + \sum_{k=0}^{p} F_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i$$

$$Y_i^{(0)} \approx \sum_{k=0}^{p} F_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i \qquad (3)$$

where:

$$Y_i^{(0)} = Y_i - f_i^{(0)}$$

The responses $Y_i^{(0)}$ are residuals. The x variables observations $F_{ik}^{(0)}$ are the partial derivatives of the mean response evaluated for each of the n cases with the parameters replaced by the starting estimates. Each regression coefficient $\beta_k^{(0)}$ represents the difference between the true regression parameter and the initial estimate of the parameter. The linear regression model approximation (3) in matrix form can be written as follows:

$$Y^{(0)} \approx F^{(0)}\beta^{(0)} + \varepsilon \qquad (4)$$

Where

$$Y_{n\times 1}^{(0)} = \begin{bmatrix} Y_1 - f_1^{(0)} \\ \vdots \\ Y_n - f_n^{(0)} \end{bmatrix},$$

$$F_{n\times(p+1)} = \begin{bmatrix} F_{10}^{(0)} & \cdots & F_{1p}^{(0)} \\ \vdots & \vdots & \vdots \\ F_{n0}^{(0)} & \cdots & F_{np}^{(0)} \end{bmatrix}$$

$$,\beta_{(p+1)\times 1}^{(0)} = \begin{bmatrix} \beta_0^{(0)} \\ \vdots \\ \beta_p^{(0)} \end{bmatrix},$$

$$\varepsilon_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Therefore, by ordinary least squares we can estimate the parameters $\beta^{(0)}$:

$$b^{(0)} = (F^{(0)'}F^{(0)})^{-1}F^{(0)'}Y^{(0)} \qquad (5)$$

Where $b^{(0)}$ is the vector of the least squares estimated regression coefficients. In matrix form, we represent the revision process as follows:

$$\theta^{(1)} = \theta^{(0)} + b^{(0)}$$

At the end of the first iteration, the revised estimated regression coefficients are $\theta^{(1)}$, and least squares criterion measure evaluated at this stage, now denoted by $SSE^{(1)}$, is :

$$SSE^{(1)} = \sum_{i=1}^{n} \left[Y_i - f(x_i, \theta^{(1)})\right]^2 = \sum_{i=1}^{n} \left(Y_i - f_i^{(1)}\right)^2$$

If the Gauss – Newton method is working effectively in the first iteration, $SSE^{(1)}$ should be smaller than $SSE^{(0)}$. The iterative process is continued until the differences between successive coefficient estimate becomes negligible. However, The Gauss – Newton method works effectively in many nonlinear regression applications. In some instances, the method may require numerous iterations before converging, and in a few cases it may not converge at all. See [1] and [2].

2.2 The M-Estimator

The "M" in the term "M-estimator" stands for maximum-likelihood-type estimator [3]. This name stems from the fact that an M-estimator can be loosely interpreted as a maximum-likelihood estimator, albeit for an unknown, non-Gaussian model. To cope with the problem of outliers, Huber [3] introduced a class of the so-called M-estimators, in which the sum of function ρ of the residuals is minimized. The estimated vector of the parameters $\hat{\beta}_M$ estimated by an M-estimator is given by

$$\hat{\beta}_M = \arg min_\beta \sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma}\right) \qquad (6)$$

A popular choice for σ is the median absolute deviation

$$\sigma = median|r_i - median(r_i)|/0.6745$$

Function $\rho(.)$ must be even, nondecreasing for positive values, and less increasing than the square. However, it is simpler to differentiate ρ with respect to β and solve for the root of the derivative. When this differentiation is applicable, the M-estimator is said to be of ψ-type. Let $\psi = \rho'$ be the derivative of ρ, and define weights such as $w_i = \psi(\frac{r_i}{\sigma})/r_i$, the estimates $\hat{\beta}_M$ is obtained by solving the system of equations:

$$\sum_{i=1}^{n} w_i^2 r_i^2 = 0$$

It is obvious that the coefficients are dependent upon the weights. Hence, to solve for M-estimators an iteratively reweighted least squares (IRLS) algorithm is employed with some initial estimates $\hat{\beta}^{(0)}$, at each iteration t until it converges. For Objective Function. Several ρ functions can be used. such as bisquare function defined as

$$\rho(z_i) = \begin{cases} \frac{c^2}{6}\left(1 - \left[1 - \left(\frac{z_i}{c}\right)^2\right]^3\right) & if\ |z_i| \le c \\ \frac{c^2}{6} & if |z_i| > c \end{cases}$$

where c is a tuning constant and $z_i = \frac{r_i}{\sigma}$. The corresponding $\psi(z_i)$ function is

$$\psi(z_i) = \begin{cases} z_i\left[1 - \left(\frac{z_i}{c}\right)^2\right]^2 & if\ |z_i| \le c \\ 0 & if |z_i| > c \end{cases}$$

The tuning constant is generally chosen to give reasonably high efficiency in normal case; in particular c = 4.685 produces a 95% efficiency when the errors are normal, while guaranteeing resistance to contamination of up to 10% of outlier. For more details, see [6].

2.3 The MM-Estimator

The "MM" in the term "MM-estimator" stands for the two stage maximum-likelihood estimator. The MM- estimator by [7] introduces the multi-stage estimator (MM-estimator). It is a combination of high breakdown and high efficiency. It can be obtained using a three-stage procedure. At first stage, an initial consistent estimate $\hat{\beta}_0$ with high breakdown point with possibly low normal efficiency is obtained. Yohai [7] suggests using the S-estimator for this stage. In the second stage, a robust M-estimator of scale parameter $\hat{\sigma}$ of the residuals based on the initial value is calculated. In the third stage, an M-estimator $\hat{\beta}$ starting at $\hat{\beta}_0$ is obtained. Huber or bi-square functions is typically used as the initial estimate $\hat{\beta}_0$. Let $\rho_0(r) = \rho_1\left(\frac{r}{k_0}\right), \rho(r) = \rho_1\left(\frac{r}{k_1}\right)$, and assume that each of the ρ_i functions is bounded, $i = 0$ and 1. The scale estimate $\hat{\sigma}$ satisfies the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{y_i - f(x_i, \theta)}{\hat{\sigma}}\right) = 0.5$$

where k_0 = 1.56 when the ρ function is biweight. In this case the estimator has the asymptotic breakdown point ($BP = 0.5$).

3. *Results and Discussion*

In this section we summarize and discus the numerical results from simulation study and real life data.

3.1 Simulations setup

Simulation is a technique for guiding the experiments of a model. Compared with analytical methods, simulation is easily understandable and highly realistic. In this study, all computations and graphics were carried out using the software package R.

3.1.1 Models

Michaelis-Menten Model

Michaelis-Menten model, used by [13] and [14], expresses the reaction velocity as a function of concentration of substrate as

$$y_i = \frac{\beta_0 x_i}{\beta_1 + x_i} + \varepsilon_i$$

Where response variable y_i is velocity and predictor variable x_i is substrate; the parameter is β_0 the maximum reaction velocity and β_1 denotes concentration of substrate. In this simulation the true parameter values are chosen to be β_0 = 5 and β_1 = 1 as in [15]. However, different true parameters are possible as long as convergence occurs in optimization process.

Exponential Model

It is a two parameter model given by the following relationship

$$y_i = \beta_0\, e^{\beta_1 x_i} + \varepsilon_i$$

Where β_0, β_1 are parameters, x is independent predictor, y response predictor, ε is random variable. Where β_0 = 0.2, β_1 = 0.3 are the true parameters (these values have been chosen arbitrary with the advantage of fast computations and the convergence occurs very quickly)

Logistic Model

It is a three parameter model given by the following relationship

$$y_i = \frac{\beta_0}{1 + e^{(\beta_1 - \log(x_i)/\beta_2)}} + \varepsilon_i$$

with β_0 = 5, β_1 = 1, β_2 = 0.6 . These values have been chosen arbitrary as in *nlr package*

3.1.2 Data generation

The contaminated normal distribution a simple useful distribution that can be used to simulate outliers [14]. For each model the explanatory x is chosen uniformly within the range (1,10). We control the outlier percentage through the outlier generating model by [14] such as:

$$\varepsilon : (1-\tau)N(\mu_1, \sigma_1^2) + (\tau)N(\mu_2, \sigma_2^2)\ .$$

Where the proportion $(1-\tau)$% refers to the percentage of non-outlier data, while the proportion τ% refers to percentage of outliers . For each sample size, 500 random data are generated with $\mu_1 = 0, \sigma_1^2 = 10, \mu_2 = 0, \sigma_2^2 = 0.2$. See appendix for how to generate data and perform simulation.

3.1.3 Comparison criterion

The mean squared errors (MSE) is used as a comparison criterion. The mean squared error is estimated by

$$MSE = \frac{\sum_{i=1}^{m}(\hat{f}_i - f)^2}{m}$$

where $\hat{f}_i$ is fitted values and f the true function with m=500 (replication number for each model).

Having examined Tables 1 to 3 carefully, we have noted many important features: The best performance has been achieved by the OLS-estimator when there are no outliers in the simulated data (0%) for all possible of sample sizes (n=50, n=100, n=150) as well as for the three different models. An interesting feature to note is that the performance of the OLS estimator is the worst one for all possible percentages of outliers (10%, 20%, 30%, 40%) as well as all possible of sample sizes (n=50, n=100, and n=150) with the three different models. For all possible of percentages of outliers (10%, 20%, 30%, 40%) the robust MM-estimator outperforms the M-estimator, and they are both are superior to the OLS estimator in all cases.

Table 1: Simulation results for first model with sample sizes (n=50,100,150) and outlier percentages (0%, 10%, 20%, 30% ,40%).

n		0%	10%	20%	30%	40%
	OLS	0.0053	0.742	0.0652	1.17	0.438
50	M	0.006	0.00287	0.00346	0.0281	0.0139
	MM	0.00571	0.00488	0.00154	0.0019	0.00689
	OLS	0.00062	0.46703	0.216	0.474	0.904
100	M	0.00105	0.00411	0.00641	0.0537	0.0211
	MM	0.00114	0.00124	0.00278	0.011	0.0119
	OLS	0.0013	0.12	0.384	0.0406	0.0812
150	M	0.00143	0.00077	0.00241	0.00035	0.00119
	MM	0.0016	0.0011	0.00056	0.00093	0.0302

Table 2: Simulation results for second model with sample sizes (n=50,100,150) and outlier percentages (0%, 10%, 20%, 30% ,40%).

n		0%	10%	20%	30%	40%
	OLS	0.00073	0.11074	4.11536	3.27919	0.39553
50	M	0.00163	0.01101	0.01088	0.05132	0.00185
	MM	0.00265	0.00563	0.00056	0.00142	0.00189
	OLS	0.00242	0.35308	0.01755	0.30446	0.33077
100	M	0.0028	0.00051	0.00117	0.0087	0.01975
	MM	0.00282	0.00091	0.00034	0.00164	0.00095
	OLS	0.00179	0.16295	0.09501	0.04144	0.17088
150	M	0.00223	0.00023	0.0016	0.00358	0.01206
	MM	0.0023	0.00051	0.00065	0.00053	0.00046

Table 3: Simulation results for third model with sample sizes (n=50,100,150) and outlier percentages (0%, 10%, 20%, 30% ,40%).

n		0%	10%	20%	30%	40%
	OLS	0.00332	0.458	0.823	3.94	2.06
50	M	0.00451	0.00172	0.232	0.347	1.34
	MM	0.00462	0.0026	0.019	0.0806	0.0344
	OLS	0.00295	0.305	0.136	0.897	1.56
100	M	0.00423	0.00346	0.00248	0.179	0.133
	MM	0.004	0.00025	0.0182	0.0648	0.165
	OLS	0.00142	1.08	0.227	0.759	1.31
150	M	0.00164	0.00415	0.0459	0.223	0.0412
	MM	0.0017	0.00175	0.0071	0.414	0.455

3.2 Real data application

The purpose of real data application is to find out whether the same conclusion drawn from simulation experiments would be in practical situations. For this, we apply the LS and the two robust estimators to Drug Concentration using Michaelis-Menten model. The source of this data is [16]. One can get the data from R package through the command data("DrugKenakin") using ***nlr*** Library. In this data, observation 5 has an outlier in the response direction. Hence, as long as there is an outlier, it could be a good example to illustrate the differences in performance of the three methods. Having examined the results in Table 4, it is obvious that the best performance is observed by MM-estimator followed by M-estimator. In other words, the MM-estimator outperforms both M-estimator and classical least square. This remark comes similar to what have been concluded in simulation results. Thus, results support the finding obtained from simulation experiments as in Figure 1.

Table 4: Parameter Estimates and mean squared error

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	MSE
OLS	7.3387	0.0641	8.351506
M	4.98951	0.04383	6.05089
MM	3.3833	0.0347	3.621733

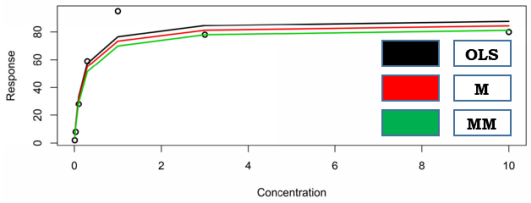

Fig. 1: Curve fitting using OLS, M, and MM for the drug concentration data.

4. Conclusion

This paper considers the problem of robust nonlinear regression in the presence of outliers. Two robust estimators are evaluated, namely M-estimator and MM-estimator using simulation study and real life data. It has been concluded that both robust estimators are superior to classical least squares with the advantages to MM-estimator in precence of outliers.

References

[1]- Bates, D. M. and Watts, D. G. (1988). Nonlinear Regression Analysis and Its Applications. New York: Wiley

[2]- Seber, G. A. F. and Wild, C. J. (1989). Nonlinear Regression. New York: Wiley.

[3]- Huber, P.H., Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35 (1964), 7-101.

[4]- Barreto, H. and Maharry, D. (2006). Least median of squares and regression through the origin. Computational Statistics & Data Analysis, 50(6), 1391-1397.

[5]- Chen, Y., Stromberg, A.J. and Zhou, M. (1997). The least trimmed squares estimate in nonlinear regression. Technical Report Department of Statistics, University of Kentucky, Lexington, KY, 40506.

[6]- Mosteller F, Tukey JW. Data analysis and regression: a second course in statistics. Addison-Wesley Ser Behav Sci Quant Methods, 1977.

[7]- Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642–656.

[8]- Hawkins, D. and Khan, D. (2009). A procedure for robust fitting in nonlinear regression.

Computational Statistics and Data Analysis, 53(12), 4500-4507.

[9]- Khalil, A., Ali, A., Khan, S., Khan, D. M. and Khalil, U. (2013). A New Efficient Redescending M-Estimator: Alamgir Redescending M-Estimator. *Research Journal of Recent Sciences*, 2(8), 79-91.

[10]- Maronna, R., Martin, R. D., & Yohai, V. (2006). *Robust Statistics Theory and Methods*. John Wiley & Sons.

[11]- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871–880.

[12]- Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis* (pp. 256–272). Springer.

[13]- Stromberg, A. and Rupert, D. (1992). Breakdown in Nonlinear Regression. *Journal of the American Statistical Association*, 87(420), 991-997.

[14]- Herwindiati D. E., Djauhar, M. a. and Mashuri, M. 2007. Robust Multivariate Outlier Labeling. Communications in Statistics - Simulation and Computation. 36(6): 1287-1294.

[15]- Tabatabai, M., Kengwoung-Keumo, J., Eby, W., Manne, U., Fouad, M. and Singh, K. (2014). A New Robust Method for Nonlinear Regression. *Journal of Biometrics & Biostatistics*, 5(5), 211.

[16]- Kenakin, TP. A Pharmacology Primer: Theory, Applications, and Methods. Third Edition. *Academic Press*; 2009. p. 286-287.

Appendix

An example of data generation using R code with model 1:

```
Set the true parameter (b0,b1), sample size (n),
and outlier percentage (p).
b0=5
b1=1
n=100
p=0.05
Generate values for  x
x=sort(runif(n, min=1,max=10))
n1=round(p*n,0)
n2=n-n1
 Generate the errors
e= c(rnorm(n2,0,0.2), rnorm(n1,0,10) )
Add the errors to true part
y=(b0*x)/(b1+x)+e
Use the data (x,y) with 5% outliers, and apply
to the three methods, then  calculate MSE
Repeat 3-5 for 500 times
Take the average of MSE
```