LOCAL ROBUSTNESS OF NONLINEAR REGRESSION M-ESTIMATORS

Shawn P. Neugebauer

Booz·Allen & Hamilton, Inc.
900 Elkridge Landing Rd.
Linthicum, Maryland 21090
spn@bah.com

Lamine Mili

Bradley Dept. of Electrical Engineering
Virginia Polytechnic Inst. & State University
Blacksburg, Virginia 24060
mililm@vt.edu

ABSTRACT

We study the local robustness properties of nonlinear regression M-estimators by analyzing their influence functions. The influence functions show that influence of position becomes, more generally, influence of model in nonlinear model estimation—indicating that we must bound not only influence of residual but also influence of model. Several examples illustrate the interpretive utility of the new influence functions. We apply the L_1 estimator to several nonlinear models and demonstrate that not only are the nonlinear regression M-estimators vulnerable to outlying observations, as in linear regression, but that non-outlying observations can cause high influence. More generally, we show that influence is caused (or mitigated) as much by model and data properties as it is by estimator properties.

1. INTRODUCTION

The goal of *regression* is to describe the structure best fitting the data. Often, a mathematical expression or *model* is available (or generated) which (possibly) explains the data. Due to the nonlinear nature of the real world, the model is often nonlinear with respect to the explanatory variables and/or the parameters. Estimation of nonlinear models, therefore, finds applications in every field of engineering and the sciences [1]. Much work has been done to build solid statistical theories on its use and interpretation [2], [3], [4], [5]. Pre-existing statistical techniques, including robust estimators, have been used directly for nonlinear model estimation problems [6], [7], [8]. However, despite this large body of nonlinear statistical theory, there has been little analysis of the tolerance of these estimators to deviations from assumptions and normality.

When estimating the parameters of a model, we desire the estimation technique to have several properties, including unbiasedness, consistency, efficiency, and robustness. *Robustness* is concerned with evaluating and improving the stability of estimation techniques when data deviate from assumptions [9], [10]. Using these concepts, the goal of *robust regression estimation* or *robust model estimation* is to estimate the model best fitting the bulk of the data. This requires techniques for measuring both the local and global robustness of an estimator. The most important tool for analyzing global robustness is the *breakdown point*, which measures the maximum fraction (ϵ) of arbitrary gross errors that an estimator can handle; the most important tool for studying local robustness is the *influence function*, which describes the effect of an infinitesimal contamination at x on an estimator T. The *maximum bias curve* is an important complement to both the influence function and the breakdown point as it links the various robustness concepts: the *gross-error sensitivity* is the slope of the tangent to the maximum bias curve at $\epsilon = 0$ and the breakdown point is the value of ϵ at which the maximum bias curve is infinite. In this paper, we focus on the local robustness properties of nonlinear regression M-estimators by analyzing their influence functions.

We define the nonlinear model estimation problem as follows. Let $\{(\underline{x}_i, y_i) : i = 1, \ldots, n\}$ be a sequence of i.i.d. random vectors such that

$$y_i = \eta(\underline{x}_i, \underline{\theta}) + e_i, \quad i = 1, \ldots, n \tag{1}$$

where $y_i \in \Re$ is the ith observation value and $\underline{x}_i \in \mathcal{X} \subset \Re^p$ is the ith row of the $n \times p$ design matrix. The variables $\underline{x}_i$ are called the *explanatory variables* or the *carriers*; the variable y_i is called the *response variable*. Let $\underline{\theta} \in \Theta \subset \Re^p$ be the p-vector of unknown parameters and $e_i \in \Re$ be the ith *error*. Finally, let $\eta(\underline{x}, \underline{\theta}) : \mathcal{X} \times \Theta \to \Re$ be the *model function*. Assume the model $\eta(\underline{x}, \underline{\theta})$ is nonlinear with respect to $\underline{x}$ and/or $\underline{\theta}$ with both observations $\{y_i\}$ and carriers $\{\underline{x}_i\}$ random. Estimates are denoted by a hat.

2. ROBUSTNESS CONCEPTS

2.1. Definition of Statistical Functionals

Let $X_1, \ldots, X_n$ be a sample from a population with distribution function F and let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic. When T_n can be written as a functional T of the empirical distribution function F_n, $T_n = T(F_n)$ where T does not depend on n, then we call T a *statistical functional.* The domain of T is assumed to contain the empirical distribution functions F_n for all $n > 1$ and the population distribution function F. The range of T is assumed to be $\Re$.

2.2. Definition of M-estimators

The specific class of statistical functionals we study are M-estimators. Let ψ be a real-valued function and let T_n be defined implicitly by

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0 \ . \tag{2}$$

The corresponding functional is defined as the solution $T(F) = \theta$ of

$$\int \psi(x, \theta) dF(x) = 0 \ . \tag{3}$$

Functionals of this form are called *M-estimators.* Common M-estimators include the L_1 estimator (or Least Absolute Deviation estimator) and the L_2 estimator (or Least Squares Estimator). M-estimators are easily extended to nonlinear regression [6],[7],[5].

2.3. Definition of the Influence Function

Following [9], [10], and [11], we say that a functional T is Gâteaux differentiable at F if there exists a linear functional L_F such that for all H

$$\lim_{t \to 0} \frac{T(G) - T(F)}{t} = L_F(H - F) \tag{4}$$

where $G = F_t = (1-t)F + tH$. L_F is called the Gâteaux derivative of T at F.

The *influence function* of T is defined as (4) when $H = \Delta_x$, the distribution with unit mass at x, yielding

$$IF(x, T, F) = \lim_{t \to 0} \frac{T((1-t)F + t\Delta_x) - T(F)}{t} \tag{5}$$

for $x \in \mathcal{X}$ where the limit exists. The influence function describes the effect of an infinitesimal contamination at x on the estimator T. We use the influence function to help identify nonlinear model estimators with unbounded influence.

3. DEFINITION OF REGRESSION M-ESTIMATORS

Given the definition of the nonlinear model estimation problem, let $K(\underline{x})$ be the distribution of $\underline{x}_i$ with density k with respect to Lebesgue measure. Assume e_i is independent of $\underline{x}_i$ and distributed according to $G(e/\sigma)$, $\sigma > 0$ with density g with respect to Lebesgue measure. Let the model distribution be defined $F_\theta(\underline{x}, y)$ with density $f_{\underline{\theta}}(\underline{x}, y) = \sigma^{-1} g((y - \eta(\underline{x}, \underline{\theta}))/\sigma) k(\underline{x})$, where $f_{\underline{\theta}}(\underline{x}, y)$ is the joint density of $(\underline{x}_i, y_i)$. We will study the model distribution $F_{\underline{\theta}}(\underline{x}, y)$ with density $f_{\underline{\theta}}(\underline{x}, y) = \phi(y - \eta(\underline{x}, \underline{\theta})) k(\underline{x})$ where $\phi(e)$ is the standard normal density. We assume a known scale σ; without loss of generality, we put $\sigma = 1$.

The M-estimator T_n in (2) is, in the more general multiparameter case, the vector solution to the system of equations

$$\sum_{i=1}^{n} \underline{\lambda}(\underline{x}_i, T_n) = \underline{0} \tag{6}$$

where $\lambda : \mathcal{X} \times \Theta \to \Re^p$. Given an empirical cumulative distribution function (c.d.f.) G_n, the functional form of the multiparameter M-estimator is $T_n = T(G_n)$, where T is defined implicitly by the vector-valued functional

$$\int \underline{\lambda}(\underline{x}, T(G)) dG(\underline{x}) = \underline{0} \ . \tag{7}$$

Regression M-estimators are a specific case of multiparameter estimation. They are defined implicitly by

$$\Gamma(T_n) = \min\{\Gamma(\underline{\theta}) : \underline{\theta} \in \Theta\} \tag{8}$$

where

$$\Gamma(\underline{\theta}) = \sum_{i=1}^{n} \rho(r) \tag{9}$$

for $\rho : \Re \to \Re^+$ and residual $r_i = y_i - \eta(\underline{x}_i, \underline{\theta})$. If the derivative of $\rho(\cdot)$ exists with respect to $\underline{\theta}$ and $\eta(\cdot)$ is twice differentiable with respect to $\underline{\theta}$ then the regression M-estimator T_n satisfies the system of equations given by

$$\sum_{i=1}^{n} \underline{\lambda}(r) = \sum_{i=1}^{n} \underline{\lambda}(\underline{x}, y, T_n) = \underline{0} \ . \tag{10}$$

Here $\underline{\lambda}(r)$ is the vector-valued function

$$\underline{\lambda}(r) = \frac{\partial}{\partial \underline{\theta}} \rho(r) = \rho'(r) \frac{\partial r}{\partial \underline{\theta}} = -\rho'(r) \frac{\partial \eta(\underline{x}, \underline{\theta})}{\partial \underline{\theta}} \ . \tag{11}$$

Put $\psi(r) := \rho'(r)$. Given an empirical c.d.f. G_n, the functional form of the regression M-estimator is $T_n = T(G_n)$, where T is defined implicitly by the vector-valued functional

$$\int \underline{\lambda}(\underline{x}, y, T(G)) dG(\underline{x}, y) = \underline{0} \ . \tag{12}$$

4. INFLUENCE FUNCTION OF NONLINEAR REGRESSION M-ESTIMATOR

Using results in [12], we write the influence function as

$$\underline{IF}(\underline{x}, y; \psi, F_{\underline{\theta}}) = \psi(r)\underline{M}^{-1}\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}} \tag{13}$$

if $\underline{M}^{-1}$ exists, where $\underline{M}$ is the $p \times p$ matrix

$$\underline{M} = \int \left\{ \psi'(r) \left[\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}}\right] \left[\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}}\right]^T - \psi(r)\frac{\partial^2\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}\partial\underline{\theta}} \right\}_{T(F_{\underline{\theta}})} dF_{\underline{\theta}}(\underline{x}, y) . \tag{14}$$

Note that when $\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}} = \underline{x}^T$, as in linear regression, (13) simplifies to

$$\underline{IF}(\underline{x}, y; \psi, F_{\underline{\theta}}) = \frac{\psi(r)}{E_K\left\{\psi'(r)\right\}} E_\Phi \left\{\underline{x}^T\underline{x}\right\}^{-1} \underline{x}^T \tag{15}$$

which is the well-known influence function for the M-estimator of linear regression. In this case, the influence function is written as a product of two factors [10],[11], specifically

$$\underline{IF}(\underline{x}, r; \psi, F_{\underline{\theta}}) = IR(r; \psi, \Phi) \cdot \underline{IP}(\underline{x}; \psi, K) \tag{16}$$

where $IR(\cdot)$ is (scalar) *influence of residual* and $\underline{IP}(\cdot)$ is (vector-valued) *influence of position*. In linear regression, *leverage points* are defined as extreme values in $\mathcal{X}$ due to their potential for unbounded influence of position.

By introducing several assumptions, we can make a similar simplification to the influence function (13). Since e and $\underline{x}$ are independent, we can write the kernel of $\underline{M}$ as a product of components due to residual r and components due to design $\underline{x}$ and model $\eta(\cdot)$. However, we are still left with a term involving the Hessian of the model at the K and G distributions. If G is symmetric and ψ is odd then this term will vanish. Alternately, we can neglect the Hessian completely. Without at least one of these assumptions, it is not possible to separate $\underline{M}$ into a product of components due to residual and components due to design and model. There is much opportunity for future work which takes into consideration the Hessian of the model.

With the stated assumptions, we write (13) as

$$\underline{IF}(\underline{x}, r; \psi, F_{\underline{\theta}}) = IR(r; \psi, \Phi) \cdot \underline{IM}(\underline{x}; \psi, K) \tag{17}$$

where $IR(\cdot)$ is (scalar) *influence of residual*

$$IR(r; \psi, \Phi) := \frac{\psi(r)}{E_\Phi\left\{\psi'(r)\right\}} \tag{18}$$

and $\underline{IM}(\cdot)$ is (vector-valued) *influence of model*

$$\underline{IM}(\underline{x}; \psi, K) := E_K \left\{ \left[\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}}\right] \left[\frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}}\right]^T \right\}^{-1} \frac{\partial\eta(\underline{x},\underline{\theta})}{\partial\underline{\theta}} . \tag{19}$$

Note that $\underline{IM}$ depends only on the model, its derivatives, and the design. It is clear that in the linear regression case, influence of model $\underline{IM}(\cdot)$ simplifies to the well-known influence of position $\underline{IP}(\cdot)$.

Assuming a bounded $\psi(\cdot)$, it is clear that $IR(\cdot)$ is bounded. However, a similar conclusion cannot be reached about $\underline{IM}(\cdot)$ without additional conditions on the model $\eta(\cdot)$. The new notation emphasizes that, for nonlinear regression M-estimators, unbounded influence is caused as much by the properties of the model as it is by the properties of the estimator and the design. The influence function can be used to gain insight on the robustness properties of M-estimators at specific models.

5. EXAMPLES

The influence function (13) is used as a tool for identifying weaknesses of design and data in M-estimators of nonlinear models. Specifically, we use the L_1 nonlinear regression estimator and calculate the estimates using the El-Attar-Vidyasagar-Dutta algorithm [8] and the Broyden-Fletcher-Goldfarb-Shanno approach [8] to unconstrained minimization.

5.1. Logarithmic Model

Consider the model

$$\eta(x_i, \underline{\theta}) = \ln\left(x_i^3 + \theta_1\right) + \theta_2$$

with design $x = \{1, 2, 3, 4, 5, 6, 10\}$. We study the exact fit properties of the M-estimators at this model with the true parameter $\tilde{\underline{\theta}} = \{0, 1\}$, that is, the estimator is studied at observed responses exactly matching the response at the known true parameter with outliers introduced for analysis purposes.

Using the starting point $\underline{\theta} = \{0.1, 0.9\}$ and the given design, we perform an L_1 estimate of the parameters of the model function. The fitted model parameters are $\hat{\underline{\theta}} = \{0.0, 1.0\}$, as expected.

Using the weakness revealed by the influence function, we calculate the Jacobian vector of the model function and find

$$\underline{J} = \frac{\partial g(x_i, \underline{\theta})}{\partial\underline{\theta}} = \begin{bmatrix} \frac{1}{x_i^3+\theta_1} \\ 1 \end{bmatrix} . \tag{20}$$

Simple calculus shows that $\lim_{x\to\infty} \frac{\partial g}{\partial \theta_1} = 0$, thus factor space outliers are not Jacobian space outliers. Since the Jacobian in the influence function of the M-estimators of nonlinear models is unbounded and directly influenced by the observations, we expect that M-estimators at this model will not be influenced by outlying x_i, as the components of the Jacobian are not unbounded.

Let $\underline{\tilde{\theta}} = \{0.0, 1.0\}$, assume exact fit, and let $y|_{x=10} = 100$. Performing an L_1 estimate of the parameters of the model function, we obtain the fitted model parameters $\underline{\hat{\theta}} = \{0.0, 1.0\}$. Perturbing the outlier further, we move it to $x = 22$ and let $y|_{x=22} = 100$. We again obtain the L_1 estimate $\underline{\hat{\theta}} = \{0.0, 1.0\}$. The perturbed data and the fitted response are shown in Figure 1. Note that if a linear model were used at this design, the outlier would have acted as a leverage point.

5.2. Bent Hyperbola Model

Consider the bent hyperbola model

$$g(x_i, \underline{\theta}) = \theta_1 + \theta_2(x_i - \theta_4) + \theta_3 \left[(x_i - \theta_4)^2 + \theta_5\right]^{1/2}$$

studied by St. Laurent and Cook [13] and Ratkowsky [2]. Using the starting point

$$\underline{\theta} = \{0.585, -0.735, -0.359, 0.062, 0.096\}$$

we perform an L_1 estimate of the parameters of the model function for the data in [2] (i.e. we do not study the exact fit properties). The data and the fitted response are shown in Figure 2.

We calculate the Jacobian vector of the model function and find

$$\underline{J} = \frac{\partial g(x_i, \underline{\theta})}{\partial \underline{\theta}} = \begin{bmatrix} 1 \\ x_i - \theta_4 \\ \left[(x_i - \theta_4)^2 + \theta_5\right]^{1/2} \\ -\theta_2 - \frac{\theta_3 (x_i - \theta_4)}{\sqrt{\theta_5 + (x_i - \theta_4)^2}} \\ \frac{\theta_3}{2\sqrt{\theta_5 + (x_i - \theta_4)^2}} \end{bmatrix} . \quad (21)$$

Application of simple calculus shows that $\lim_{x\to\infty} \frac{\partial g}{\partial \theta_2} = \infty$, $\lim_{x\to\infty} \frac{\partial g}{\partial \theta_3} = \infty$, $\lim_{x\to -\infty} \frac{\partial g}{\partial \theta_2} = -\infty$, and $\lim_{x\to -\infty} \frac{\partial g}{\partial \theta_3} = \infty$. Since the Jacobian in the influence function of the M-estimators of nonlinear models is unbounded, and these components of the Jacobian of this model are unbounded in x, we expect that M-estimators at this model will be highly influenced by outlying x_i.

To confirm our expectations, we perturb the x_i component of a single observation such that the observation becomes an outlier in the factor space X. Specifically, we move the first observation from $(-1.39, 1.12)$ to $(-10, 1.12)$. Performing an L_1 estimate of the parameters of the model function, we obtain the fitted model parameters

$$\underline{\hat{\theta}} = \{1.5094, -0.7563, -0.7880, 0.07520, 1.7910\} .$$

The perturbed data and the fitted response are shown in Figure 3. We observe that the fitted response follows the outlier exactly. This confirms our expectation of the low tolerance of the nonlinear regression M-estimators to factor space outliers at this model. In this simple case, we might have flagged the perturbed data as a potential leverage point for no other reason than it being an outlier in X.

Further examples demonstrate other, dangerous, highly influential behavior not directly revealed by the influence function and a source of current research. To illustrate the low tolerance of the nonlinear regression M-estimators to non-outlying, vertically perturbed observations, we move the first two observations from $(-1.39, 1.12)$ to $(-1.39, 0.9)$. The perturbed data and fitted response are shown in Figure 4. We note that the fitted response follows the perturbation *exactly*—these observations are *highly* influential at this model despite their non-outlying nature. Traditional techniques would not identify these values. The new influence function gives us a tool for identifying the weaknesses of design and data in nonlinear model estimation.

6. REFERENCES

[1] S.A. Kassam and H.V. Poor. Robust techniques for signal processing: a survey. *Proceedings of the IEEE*, 73:433–481, 1985.

[2] David A. Ratkowsky. *Nonlinear Regression Modeling: A Unified Practical Approach.* Marcel Dekker, 1983.

[3] A. Ronald Gallant. *Nonlinear Statistical Models.* John Wiley & Sons, New York, New York, 1987.

[4] Douglas M Bates and Donald G Watts. *Nonlinear Regression Analysis and its Applications.* John Wiley & Sons, New York, New York, 1988.

[5] G.A.F. Seber and C.J. Wild. *Nonlinear Regression.* John Wiley & Sons, New York, New York, 1988.

[6] R. Dutter and P.J. Huber. Numerical methods for the nonlinear robust regression problem. *J. Stat. Comput. Simulation*, 13:79–114, 1981.

[7] Gao Li and Kaj Madsen. Robust nonlinear data fitting. In *Numerical Analysis 1987 (Dundee, 1987)*, Pittman Research Notes Math. Series, 170, pages 176–191. Longman Sci. Tech., New York, New York, 1988.

[8] René Gonin and Arthur H. Money. *Nonlinear L_p-norm Estimation.* Marcel-Dekker, New York, New York, 1989.

[9] Peter J. Huber. *Robust Statistics.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, New York, 1981.

[10] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, New York, New York, 1986.

[11] Frank R. Hampel. The influence curve and its role in robust estimation. *JASA: Theory and Methods*, 69(346):383–393, June 1974.

[12] Shawn P. Neugebauer. Robust analysis of M-estimators of nonlinear models. Master's thesis, Virginia Polytechnic Institute & State University, Blacksburg, Virginia, August 1996.

[13] Roy T. St. Laurent and R. Dennis Cook. Leverage and superleverage in nonlinear regression. *JASA, Theory and Methods*, 87(420):985–990, December 1992.

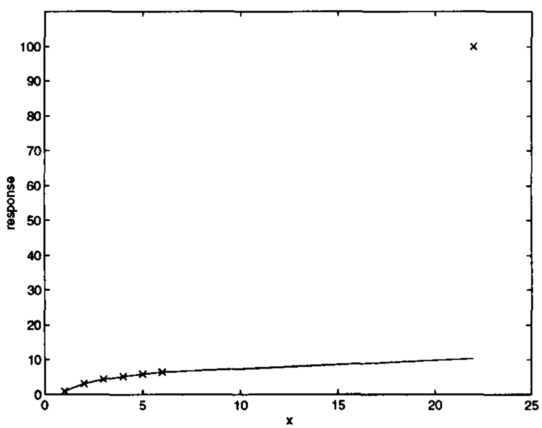

Figure 1: Fit of Logarithmic Model at True Data with Large x_i and y_i Perturbation

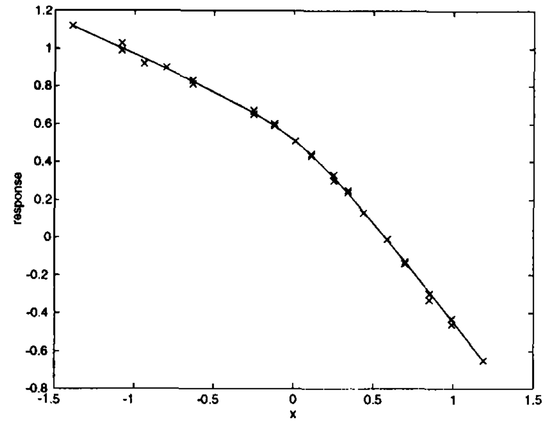

Figure 2: Fit of Bent-Hyperbola Model to Band Height Data

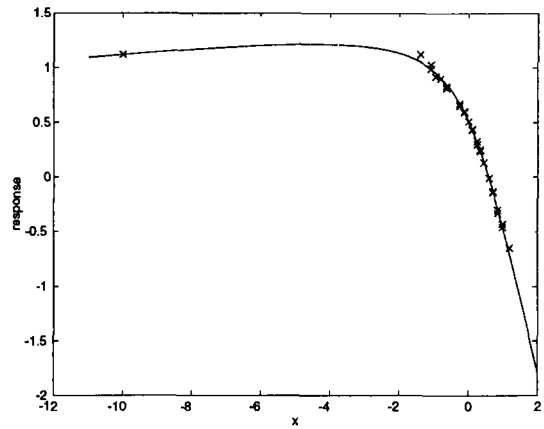

Figure 3: Fit of Bent-Hyperbola Model to Band Height Data with Single Large x_i Perturbation

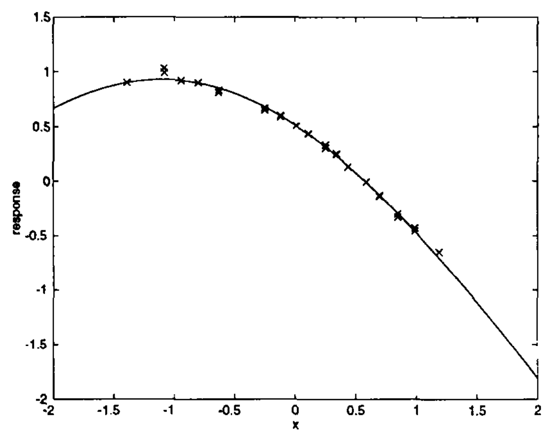

Figure 4: Fit of Bent-Hyperbola Model to Band Height Data with Two Non-outlying Vertical Perturbations