Is this the least squares estimate?

By EUGENE DEMIDENKO

Dartmouth Medical School and Dartmouth College
7927 Rubin Building, One Medical Center Drive, Lebanon, NH 03756, USA
e-mail: eugene.demidenko@dartmouth.edu

To be published in Biometrika

Summary

It is shown that the sum of squares can have several local minima with a positive probability for any intrinsically nonlinear regression with in¯nite tails. Therefore, the availability of global criteria is crucial. The concept of the local convexity level for sum of squares in nonlinear regression model is introduced. A general formula for the local convexity level of the sum of squares is derived and link to the curvature of nonlinear model is established. It is shown that the local convexity level is equal to the minimum of the squared radius of the full curvature of the expectation surface of nonlinear regression. Two general global criteria are formulated. The calculation of the local convexity level and construction of global criteria are illustrated by four types of nonlinear models: polylinear, power, linear M-regression, and exponential regression models. The suggested global criteria work well for real life data.

Some key words: nonlinear regression; uniqueness; convexity; unimodality; robust regression; curvature.

1. Introduction

Only in rare cases estimation procedure admits a closed form solution, the least squares estimate in linear regression model is such a fortunate example. Then, after iterations on minimization of the sum of squares in nonlinear regression are done and we look at the ¯nal parameter value, a reasonable question arises: "Is this the least squares estimate?", i.e. did we ¯nd the global minimum? Amazingly, until today it is extremely di±cult to answer this simple to formulate question. Moreover, as is shown in this paper, it is quite possibly for a sum of squares to have at least two local minima.

The global optimization is under constant interest, particularly a special journal has been established recently, Journal of Global Optimization. Several books have been published on the topic. The objective of the global optimization is to ¯nd the global minimum of

a continuous function (possibly under some restrictions). Two approaches are available: deterministic and stochastic. In deterministic approach the parameter set is divided into a ¯nite number of small subsets where it is known there is a unique local minimum, e.g. Horst and Tuy (1996), Floudas and Pardalos (1992), Hansen (1992). The drawbacks of this approach is that it requires a compact parameter set and the number of the divided subsets may be very large. In stochastic global search the global minimum is found with a certain probability, Mockus (1989), Zhigljavsky (1991). In several books merits of di®erent global optimization techniques are compared, Horst and Pardalos (1995), Kearfort (1996), Pinter (1996). The lack of well working global criteria could be explained by the fact that all general criteria, i.e. applied to a general class of optimization functions cannot work because they are not speci¯c, Horst and Tuy (1996). The sucsess in construction of a global criterion is determined by the class of nonconvex functions it covers. The sum of squares for nonlinear regression is a very interesting and broad class of multiextremal functions yet allowing construction of working global criteria.

We argue that in practical optimization, particularly in optimization driven by statistical problems we can rarely expect that the minimized function has many local mimima. Thus there is just a possibility of multiextremality. Consequently, instead of doing a very expensive global search we need just criteria for global minimum, i.e. conditions under which a found local minimum is in fact the global one. To work out such criteria is the goal of the present paper. In particular, links between the full curvature of the nonlinear model and local convexity of the sum of squares are established: the minimum of the squared radius of the full curvature of the expecation surface is equal to the level of the local convexity of the sum of squares. Based on the introduced concepts we are able to formulate global criteria, i.e., to propose conditions under which a found local minimizer is the global one.

Now let us review some computational basics of nonlinear regression estimation and give necessary de¯nitions. Let y_i be the ith observation and $f(x_i;®)$ the regression function where x_i is a ¯xed vector, $®2 £½R^m$ is a unknown vector parameter. We assume that f is a twice continuously di®erentiable function and $£$ is a convex $m ¡$ dimensional parameter set. Since fx_ig are nonrandom we can simplify the notation letting $f_i(®) = f_i(x_i;®)$. Then, the standard nonlinear regression model takes the form:

$$y_i = f_i(®)+²_i; \qquad i = 1;:::;n \tag{1.1}$$

where it is assumed $E(²_i) = 0$; $f²_1;:::;²_ng$ are independent and identically distributed with a positive probability function, and $n ¸ m$. The set of points $ff_1(®);::;f_n(®)g$ in R^n is called expectation surface (e.g. Bates and Watts 1988, Seber and Wild 1989). The object of our

study is the Sum of Squares (SS), as a function of parameter given data,

$$S(\theta) = \sum_{i=1}^{n} (y_i - f_i(\theta))^2. \qquad (1.2)$$

Our concern is the global minimum of $S(\theta)$ on Θ. The point of the minimum is called the Least Squares Estimate (LSE). Geometrically, the LSE corresponds to the nearest point on the expectation surface from the data vector $y=(y_1,\ldots,y_n) \in R^n$. It is assumed that the nonlinear regression is regular (e.g., Gallant 1987, Pazman 1993), i.e. the derivative matrix has full rank, $\mathrm{rank}(G(\theta)) = m$ for all $\theta \in \Theta$; where $G(\theta)$ is the $n \times m$ matrix of first derivatives, $\{\partial f/\partial\theta_1,\ldots,\partial f/\partial\theta_m\}$; $f = (f_1,\ldots,f_n)'$. For linear model it means that the design matrix has full rank. In most cases when the derivative matrix has not full rank the nonlinear regression model is not identified, i.e., for different θ_1 and θ_2 we have $f(\theta_1) = f(\theta_2)$.

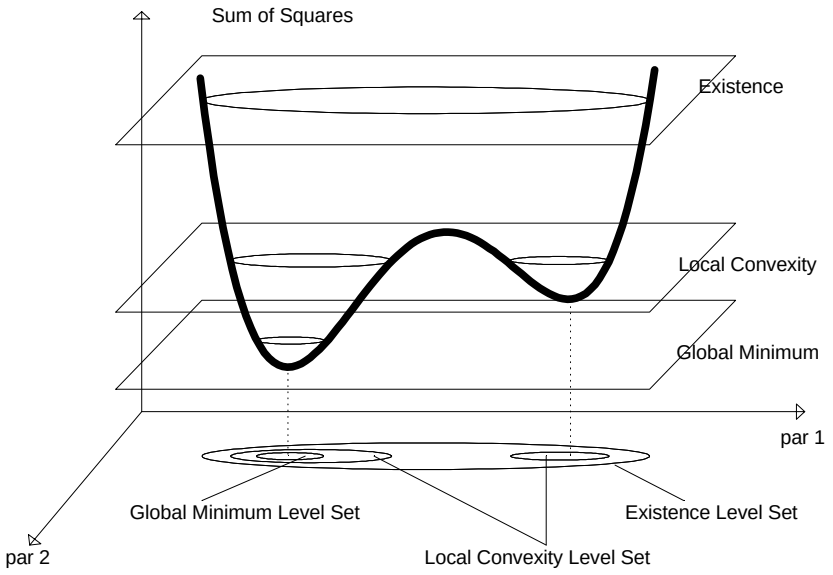

Fig. 1. Three level sets of the Sum of Squares: the Existence LS is compact; the Local Unimodality LS consists of two components corresponding to two local minima with only one local minimizer on each component. The Local Convexity LS is convex, and the SS is a convex function on this set.

In this paper we investigate the properties of the sum of squares in terms of its convexity and possibility to have several local minima, refer Fig.1 for geometrical illustration. The

concept of the Level (or Lebegue) Set (LS),

$$L(S_¤) = f® 2 £: S(®) < S_¤ g; \tag{1.3}$$

given level $S_¤$; is crucial to our analysis. Our general strategy is to express properties of the SS through its levels. For instance, the Existence Level (EL), as the minimum of the SS on the border of the parameter set, was de¯ned in Demidenko (1981, 1989, 1996), Nakamura (1984). If the parameter set £ coincides with the entire space the EL is de¯ned as

$$\overline{S}_E = \lim_{r \to 1} \inf_{k®k > r} S(®):$$

Then the existence criteria works simply: if there is an initial parameter $®_0$ such that the value of the SS is less than EL, $\overline{S}_E$ then the LSE exists. It is easy to prove that the level set $L(S(®_0))$ is compact, implying that any minimization algorithm produces a sequence of parameters with at least one limit point, Demidenko (1996). In the present paper we introduce another level set, the Local Convexity Level (LCL) set, as the level set where the SS is locally convex, i.e., the Hessian is positive de¯nite. Based on this level we can construct ultimate global criteria. Again, the criterion works easy: if the value of the found local minimum of the SS is less than the calculated threshold value, then this minimum is the global one, see Fig.1.

The outline of the paper is as follows. We start with the warning result that for an intrinsically nonlinear regression with in¯nite tails the SS quite likely may have at least two local minima, i.e. to be multimodal. In the third section the LCL is introduced and simple global criteria based on general optimization theory are formulated. The following sections contain several examples of the LCL calculation and global criteria construction: polylinear and power regression, M-regression, exponential regression.

2. How likely can one expect multimodality of the Sum of Squares?

The possibility to have several local extrema is studied here for the Sum of Squares in nonlinear regression model. How likely the Sum of Squares (SS) may have two or more local minima? In particular, do there exist classes of nonlinear regressions with unimodal SS for any data y? As it is shown in this section, for any intrinsic nonlinear regression with in¯nite tails, the probability that the SS has at least two local minima is positive. Consequently, there are no nonlinear regressions with unimodal SS for any data. Therefore, the problem of determining whether a found local minimizer is indeed the global one, i.e., the LSE, is crucial.

In this section it is assumed that the parameter set coincides with the entire space R^m: The following result on multimodality will be proved for regressions with in¯nite tails: $k®k ! 1$

implies $k\, f(\theta)\, k \to \infty$. We say that a multivariate regression is intrinsically linear if its regression function can be written as $f(\theta) = g_1(\theta)x_1 + \ldots + g_m(\theta)x_m + z$ where $x_1, \ldots, x_m, z \in R^n$; the system of vectors $\{x_1, \ldots, x_m\}$ has full rank, and $g_1(\theta), \ldots, g_m(\theta)$ are twice continuously differentiable scalar functions such that the map $g: R^m \to R^m$ is injective (Pazman 1993). A characteristic feature of intrinsically linear regression is that its expectation surface lies in a linear $m-$dimensional manifold. We call regression intrinsically nonlinear if it is not intrinsically linear. Now we formulate the main result on multimodality of the SS for nonlinear regression.

Theorem 2.1. The probability, that the Sum of Squares for any three times differentiable intrinsically nonlinear regression with infinite tails and $n \geq m(m+3)/2$ has at least two local minima, is positive.

Proof of this Theorem is found in the Appendix.

On contrary, as was shown by Pazman (1984), the probability that several local minimizers give the same value for the global minimum of the SS, is zero.

Example. The concept of possible multimodality of the SS is illustrated on the one-parameter quadratic regression model of the form $y_i = \theta x_{i1} + \theta^2 x_{i2} + \varepsilon_i$ where $-1 < \theta < 1; \sum x_{i2}^2 > 0$: It is easy to show that this regression has infinite tails. We also assume that $\{x_{i1}\}$ and $\{x_{i2}\}$ are not collinear which implies that this regression is intrinsically nonlinear. Therefore, the above theorem is applicable and the probability to have at least two local minimum of the SS is positive. We aim to determine regions in the outcome space where the SS is not convex, multiextremal, has at least two local minima, etc. For simplicity I consider the regression on the plane with orthogonal covariates, i.e. $n = 2$; and $x_{11} = 1; x_{12} = 0$; and $x_{21} = 0; x_{22} = 1$: Then $f_1(\theta) = \theta$; $f_2(\theta) = \theta^2$ and the expectation curve becomes a parabola in R^2 (Fig. 2). We have

$$S(\theta) = (y_1 - \theta)^2 + (y_2 - \theta^2)^2; \quad \frac{1}{4}\frac{dS}{d\theta} = \theta^3 + \theta(\frac{1}{2} - y_2) - \frac{1}{2}y_1; \quad \frac{1}{4}\frac{d^2S}{d\theta^2} = 3\theta^2 + (\frac{1}{2} - y_2):$$

Therefore, the region of the convexity of the SS is below the line $y_2 = 1/2$: This means that if $y_2 < 1/2$ for the data vector $(y_1; y_2)$ then $d^2S/d\theta^2 > 0$ for all $\theta \in R^1$: If $y_2 > 1/2$; there exists at least one θ_0 such that $d^2S/d\theta^2(\theta = \theta_0)/d\theta^2 < 0$: Now we find the region of multimodality. Since $dS/d\theta = 0$ is a cubic equation for θ we can apply the condition for the existence of three different real roots (e.g., Abramowitz and Stegun p.17, 1972) which leads to the inequality $y_2 > 3/\sqrt[3]{4}\, |y_1|^{2/3} + 1/2; \; -1 < y_1 < 1$: Therefore, if $(y_1; y_2)$ satisfies the preceding inequality, the SS will have two local minima.

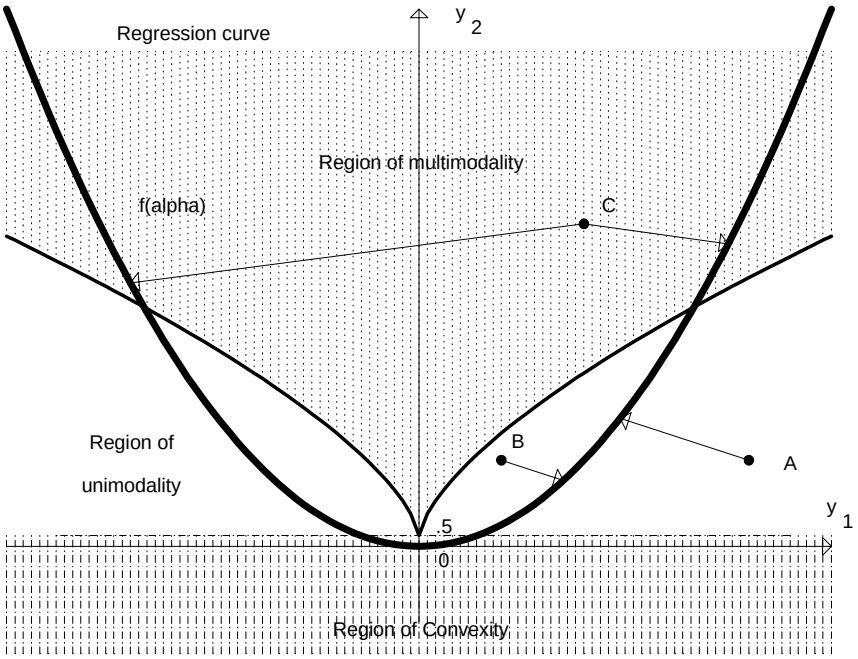

Fig. 2. The regions of the local convexity and multimodality of the SS for the quadratic one-parameter regression. Bold line { regression expectation curve $y_2 = y_1^2$ (parabola). For points A and B there is only one local minimum of the SS, for point C two local minimizers exist, the true least squares estimate corresponds to the right arrow because the distance to the curve is smaller. The region of the local (global) convexity is below the line $y_2 = :5$:

Three regions are shown in Fig. 2. Here, for data vectors A and B the according SS has one local minimum although it is not a convex function (arrows point to the nearest point on the expectation curve, i.e., the LSE). On contrary, for the observation point C the normal equation dS=d® = 0 has two solutions, and the LSE corresponds to the right, shortest arrow. For given ¾2 and the true parameter ® we can calculate

$$\mathrm{prf SS\ has\ two\ local\ minimag} = \frac{1}{\frac{1}{4}\frac{3}{4}^2} \int_0^{\tilde{A}} \int_{3=3\sqrt{4y_1^{2=3}+0:5}}^{1} e^{\,i\,0:5(y_2\,i\,\circledR^2)=\frac{3}{4}^2} dy_2 \Bigg)^{!} e^{\,i\,:05(y_1\,i\,\circledR)=\frac{3}{4}^2} dy_1:$$

Let us take for example ¾2 = 1: Then, for ® = 0 the probability of multimodality is quite small, 0:056. However, for moderate ® = 1 it is 0:15 and for large ® = 3 the probability that the SS for the quadratic model has two local minima is 0:999:

3. Local convexity

The concept of convexity is fundamental in the theory of optimization, e.g., Rockafellar (1970), Ortega and Rheinboldt (1970), Webster (1994), Chong and Zak (1996). For the

reader's convenience I formulate essential concepts of the optimization theory here, in terms of an arbitrary function.

It is said that a set $Q \subset R^m$ is convex if along with any pair of points it contains the whole segment. Let $F(u)$ be an arbitrary continuous function of vector argument $u \in Q \subset R^m$ where Q is a convex m-dimensional set, or function domain. We say that $F(u)$ is convex on Q if for all $u_1, u_2 \in Q; u_1 \neq u_2; 0 < \alpha < 1$ we have $F(\alpha u_1 + (1 - \alpha)u_2) \leq \alpha F(u_1) + (1 - \alpha)F(u_2)$: It is said that $F(u)$ is strictly convex if the last inequality is strict. We say that u_* is the global minimizer, or the point of the global minimum on Q if $F(u_*) < F(u)$ for all $u \in Q; u \neq u_*$: As follows from this definition, the global minimizer is unique, if it exists. It is said that u_* is a local minimizer, or a point of local minimum if there is a neighborhood V at u_* such that $F(u_*) < F(u)$ for all $u \in Q \cap V; u \neq u_*$: We say that $F(u)$ is unimodal if it has only one local minimum, otherwise we call it multimodal. The central theorem of the optimization theory is that a strictly convex function on a convex set has at most one local minimizer which is the global one. We say that F is locally strictly convex on a set if it is locally strictly convex at each point of the set. It is easy to show that a locally strictly convex function on a convex set is convex. There is a well known sufficient criterion for the local convexity: $F(u)$ is strictly locally convex at u_* if its Hessian matrix at u_* is positive definite (p.d.). These facts are enough to formulate the following

General criterion for the global minimum, I (GCGMI). Let $F(u)$ be a function on $Q \subset R^m$ which is locally strictly convex on a convex subset $L \subset Q$. Let us define $\underline{F}_{L^c} = \inf_{u \in L} F(u)$: If $u_* \in Q$ is a local minimizer and $F(u_*) < \underline{F}_{L^c}$ then u_* is the global minimizer of F on Q:

Proof is short yet important for a better understanding of how this criterion works. It is easy to see that $u_* \in L$ because otherwise $F(u_*) \geq \underline{F}_{L^c}$: Now let u be any point from Q: If $u \in L$ and $u \neq u_*$ we have $F(u) > F(u_*)$ because F is locally convex on a convex set. If $u \notin L$ then $F(u_*) < \underline{F}_{L^c} \leq F(u)$: Therefore, u_* is the global minimizer on Q.

We have to mention that this criterion, as all others considered later, is sufficient, not necessary. In other words, its failure does not imply that the minimized function is multimodal.

Now we introduce an important concept, the local convexity level. We know that at the minimum the Hessian is at least nonnegative definite. Clearly, lower the value of the function is more likely the Hessian is positive definite. Therefore, we can define a region where the Hessian is positive definite as a collection of parameter points which give the value of the function less than a certain value (level). It explains the following definition.

Definition 3.1. A value F_{LC} is called Local Convexity (LC) level for function $F(u)$ if

its Hessian is positive definite on the level set $L(F_{LC}) = \{x : F(x) < F_{LC}\}$.

We can speak of the Upper LC level as the maximum of all LC levels. The concept of the LC level provides the following global criterion.

General criterion for the global minimum, II (GCGMII). If the level set $L(F_{LC})$ is convex and a local minimizer u_* belongs to it, then u_* is the global minimizer.

Proof is again easy but worth to present. Since $F(u)$ is a convex function on a convex set $F(S_{LC})$ we obtain $F(u_*) < F(u)$ for all $u \in L(S_{LC})$; $u \neq u_*$. If $u \in L(F_{LC})$; by definition $F(u) \geq F_{LC} > F(u_*)$. Thus, u_* is the global minimizer.

In the rest of this section we apply this theory to the Sum of Squares (1.2). Notice, the SS is locally strictly convex at θ if matrix

$$\frac{1}{2}\frac{\partial^2 S}{\partial \theta^2} = G'(\theta)G(\theta) - \sum_{i=1}^{n} (y_i - f_i(\theta))H_i(\theta) \tag{3.1}$$

is p.d. where $G(\theta) = \partial f/\partial \theta$ is the $n \times m$ matrix of the first derivatives and $H_i(\theta) = \partial^2 f_i/\partial \theta^2$ is the $m \times m$ symmetric matrix of the second derivatives of f_i, $i = 1, \ldots, n$.

We want to express the positive definiteness of (3.1) in terms of the SS-value. For this purpose let us multiply (3.1) by p' and p where p is an arbitrary nonzero m-vector ($'$ denotes vector/matrix transpose, hereafter). Applying the Cauchy inequality to the second term, $\sum (y_i - f_i)p'H_i p$ we obtain

$$p'\frac{1}{2}\frac{\partial^2 S}{\partial \theta^2}p \geq p'G'(\theta)G(\theta)p - \sqrt{S(\theta)}\sqrt{\sum_{i=1}^{n} (p'H_i(\theta)p)^2}. \tag{3.2}$$

Define the radius of the full curvature as a function of the parameter vector θ and the $m \times 1$ direction vector p

$$R_F(\theta;p) = \frac{p'G'(\theta)G(\theta)p}{\left|\sum_i (p'H_i(\theta)p)^2\right|^{1/2}}; \tag{3.3}$$

following Bates and Watts (1988). If the derivative matrix G has full rank the numerator is positive, thus we can put $R_F = +\infty$ if the denominator is zero. Thus, based on the inequality (3.2) we can determine the LC level in nonlinear regression as follows.

Local Convexity Level for SS. The Upper Local Convexity Level (ULCL) for the SS of nonlinear regression model (1.1) is equal to the minimum of the squared radius of the full curvature of the expectation surface, or

$$\overline{S}_{LC} = \min_{\theta} \min_{p} R_F^2(\theta;p). \tag{3.4}$$

The Local Convexity Level (LCL) is any number less or equal $\overline{S}_{LC}$; and will be denoted S_{LC}.

Shortly, the ULCL is the minimum of the squared radius of the full curvature. The word "upper" in the last de¯nition can be justi¯ed by the following comment. If for some $®_0$ and p_0 we have $R_F^2(®_0;p_0) > \overline{S}_{LC}$ then there are observations $y_1;:::;y_n$ with the $SS = \sum (y_i ¡ f_i(®_0))^2 = R_F^2(®_0;p_0)$ and $p_0^0(@^2S(®_0)=@®^2)p_0 < 0$; i.e. the Hessian is not p.d. at $®=®_0$: It follows from the fact that the Cauchy inequality in (3.2) turns into equality if vectors $(y_1 ¡ f_1;:::;y_n ¡ f_n)$ and $(p_0^0H_1p_0;:::;p_0^0H_np_0)$ are collinear.

Now we will consider the problem of calculating $\min_p R_F^2(®;p)$ given ® (therefore ® is omitted). As follows from (3.3) this optimization problem is equivalent to

$$' = '(p;GjH_i; i = 1;::;n) = \frac{(p^0G^0Gp)^2}{\sum (p^0H_ip)^2} \;) \min_p : \tag{3.5}$$

We show that this problem can be reduced to a series of eigenvalue problems; called repeated eigenvalue problem. Indeed, let T be the Cholesky decomposition factor of G^0G; i.e., T is a triangle nonsingular matrix such that $T^0T = G^0G$: Denote $q = Tp$; then $p^0H_ip = q^0Q_iq$ where $Q_i = (T^0)^{i\,1}H_iT^{i\,1}$ and $'_{min}$ is equal to reciprocal of

$$'_{max} = \max_{q^0q=1} \sum (q^0Q_iq)^2: \tag{3.6}$$

To solve (3.6) we denote $K(q) = \sum Q_iqq^0Q_i$; then $\sum (q^0Q_iq)^2 = q^0K(q)q$: Let q_0 be any initial vector of unit length. We solve the eigenvalue problem $q^0K(q_0)q \;) \max$ and obtain the solution q_1: Then matrix K is recalculated, i.e., we solve $q^0K(q_1)q$ and repeat this process until convergence. A good starting value for q would be the eigenvector correspondent to the maximum eigenvalue of matrix $\sum Q_i^2$: The sequence fq_sg generated by this process has at least one limit point because it belongs to a compact set, the sphere of the unit radius. Moreover, the sequence $fq_s^0K(q_s)q_sg$ is increasing because

$$q_s^0K(q_s)q_s \cdot q_{s+1}^0K(q_s)q_{s+1} = q_s^0K(q_{s+1})q_s \cdot q_{s+1}^0K(q_{s+1})q_{s+1}; \qquad s = 0;1;:::$$

The minimum of (3.5) is denoted as $'_{min} = 1='_{max}$: Since $q^0q = 1$ it implies $(q^0Q_iq)^2 \cdot p^0H_i^2p$; and $'_{min} ¸ ¸_{min}(G^0G(\sum H_i^2)^{i\,1}G^0G)$: Obviously, for linear regression $\sum H_i^2(®) = 0$ and $\overline{S}_{LC} = 1$:

It is important to understand that it is not required calculating the value $\overline{S}_{LC}$ precisely using some optimization algorithm (otherwise all above does not make sense because it might be even more di±cult task than the SS minimization itself). Instead, we have to conduct an analytical investigation to ¯nd a lower bound for $\overline{S}_{LC}$; the examples can be found in the next section.

4. Examples. General

In the following sections the calculation of the local convexity level and construction of global criteria are illustrated by four examples of nonlinear regression model. In the first example the exact value $\overline{S}_{LC}$ is calculated. In the second example the LCL is derived using some specific inequalities. In the third example we demonstrate how the concept of the LCL can be generalized to functions different from the SS, and a global criterion is formulated based on the general criterion for the global minimum formulated in the previous section. At last, the exponential models are considered: the formula for the LCL is provided and a global criterion is suggested. Almost all criteria use the repeated eigenvalue problem at the final stage, considered in the preceding section.

As the reader can realize, every type of nonlinear model requires some creativity in finding a lower bound for the Hessian of the SS and constructing a global criterion. In particular, at some point one has to apply certain inequalities which take into account specific structure of the considered nonlinear regression model.

5. Polylinear regression

This type of nonlinear regression is defined by the function

$$f(\theta;\rho) = X\theta + \rho u - \rho Z\theta; \qquad \theta \in R^m; \rho \in R^1 \tag{5.1}$$

where X and Z are fixed $n \times m$ design matrices, matrix Z has full rank, $m < n$; and u is a $n \times 1$ vector. This kind of model emerges in time series regression analysis where residuals follow the first-order autoregression process (e.g., Judge et al. 1982). More precisely, if $y_t = \theta' x_t + \varepsilon_t$ is the original linear model and $\varepsilon_t = \rho\varepsilon_{t-1} + \delta_t$ where δ_t is the uncorrelated error term with zero mean and constant variance, we can apply first-order differences with unknown coefficient ρ to remove the existing correlation between ε_t and ε_{t-1}: This gives us a nonlinear regression $y_t - \rho y_{t-1} = \theta' x_t - \rho\theta' x_{t-1} + \delta_{t-1}$ which has a particular form (5.1).

For regression (5.1) we have $G = \partial f/\partial(\theta;\rho) = [X - \rho Z; u - Z\theta]$; the $n \times (m+1)$ matrix, and

$$\frac{1}{2} p' \partial^2 S/\partial\gamma^2 p = p' G'(\gamma) G(\gamma) p - \sum (y_i - f_i(\gamma)) p' H_i p$$

where

$$H_i = \begin{bmatrix} 0 & z_i \\ z_i' & 0 \end{bmatrix}; \tag{5.2}$$

and z_i is the ith row of matrix Z; $\gamma = (\theta';\rho)'$: We notice that the denominator of (3.3) for this model does not depend on parameter, and consequently in order to find the minimum over γ it suffices to consider only the numerator. Let $p = (p_1';p_2)'$ be fixed, then we have

$$\min_{\gamma} p' G'(\gamma) G(\gamma) p = \min_{\gamma} \| Gp\|^2 = \min_{\theta;\rho} \| (X - \rho Z)p_1 + (u - Z\theta)p_2\|^2$$

$$= \min_{\alpha,\beta} \| (Xp_1 + p_2 u) - Z(p_2\alpha + \beta p_1)\|^2 .$$

Since α and β are unrestricted vector $\zeta = \alpha p_2 + p_1\beta$ covers the entire space R^m, so that we may write

$$\min_{\alpha,\beta} \| (Xp_1 + p_2 u) - Z(p_2\alpha + \beta p_1)\|^2 = \min_{\zeta} \| Wp - Z\zeta\|^2 ,$$

where $W = [X, u]$. However, for fixed Wp the minimum over ζ can be found via least squares:

$$\min_{\theta} p' G'(\theta) G(\theta) p = p' W' (I - Z(Z'Z)^{-1} Z') W p .$$

Hence, finally

$$\bar{S}_{LC} = \min_{p} \frac{(p' W' (I - Z(Z'Z)^{-1} Z') W p)^2}{\sum (p' H_i p)^2} . \qquad (5.3)$$

Finally, we notice that the minimum in (5.3) may be found by solving the repeated eigenvalue problem considered previously. Therefore, for the polylinear regression (5.1), the Hessian is positive definite if sum of squares less than $\bar{S}_{LC}$. Notice that $\bar{S}_{LC} > 0$ if and only if any linear combination of vector-columns of matrix $[X, u]$ cannot be expressed as a linear combination of vector-columns of matrix Z; e.g., matrix $[X, u, Z]$ has full rank $2m + 1 \le n$.

6. Power regression

We define power regression as a regression with function $f_i(\theta) = (\theta' x_i)^p$, where $p \neq 0$ is given, and $x_1, \ldots, x_n$ are linearly independent. $p = 1$ gives standard linear regression model, $p = -1$ corresponds to hyperbolic, or Michaelis-Menten model (Bates and Watts 1988). Power regressions with values $p = \pm 1/2$ are also popular (e.g., Ezekiel and Fox 1959). The parameter set is defined as

$$\Theta = \{\theta \in R^m : \theta' x_1 \ge 0, \ldots, \theta' x_n \ge 0\} \quad \text{when} \quad p > 0;$$
$$\Theta = \{\theta \in R^m : \theta' x_1 > 0, \ldots, \theta' x_n > 0\} \quad \text{when} \quad p < 0.$$

As follows from its definition the parameter set is convex; also it will be assumed that it is not empty and has dimension m. For the sake of simplicity the following development assumes $y_i > 0, i = 1, \ldots, n$; a similar analysis can be accomplished in case when some of y_i are negative.

For the power regression

$$\frac{1}{2p} \frac{\partial^2 S}{\partial \theta^2} = X' D(\theta) X \qquad (6.1)$$

where the $n \times m$ matrix X consists of vectors $\{x_i\}$; and $D(\theta)$ is the $n \times n$ diagonal matrix with diagonal elements

$$D_{ii}(\theta) = (2p - 1) d_i^{2(p-1)} - (p - 1) y_i d_i^{p-2} = p(y_i - e_i)^{2 - \frac{2}{p}} - (p - 1) e_i (y_i - e_i)^{1 - \frac{2}{p}} \qquad (6.2)$$

where $d_i = \beta' x_i$ and $e_i = y_i - d_i^p$ is the ith residual.

We aim to find local convexity level, i.e. level sets of SS where the Hessian, (6.1) is positive definite (diagonal elements of matrix D are positive). The analysis of positive definiteness of the Hessian depends on the value of p: We consider some special cases:

1. $1/2 \le p \le 1$: Since $(p-1)y_i \le 0$ and $(2p-1)d_i^p \ge 0$ the SS is a strictly convex function on $\mathcal{L}$. Consequently, the SS is unimodal, and any local minimizer is the global one.

2. $p > 2$: The main idea is to find lower bound for (6.2) in terms of a linear function of the residual, e_i: For the first term of the right hand side of (6.2), applying the inequality $(1-x)^v \ge 1 - vx; v \ge 1; x \le 1$; we obtain

$$(y_i - e_i)^{2-\frac{2}{p}} = y_i^{2-\frac{2}{p}}\left(1 - \frac{e_i}{y_i}\right)^{2-\frac{2}{p}} \ge y_i^{2-\frac{2}{p}}\left(1 - \left(2-\frac{2}{p}\right)\frac{e_i}{y_i}\right) = y_i^{2-\frac{2}{p}} - \frac{2(p-1)}{p}e_i y_i^{1-\frac{2}{p}}.$$

For the second term of (6.2) we apply another elementary inequality:

$$-x(y-x)^\alpha \ge -\frac{1}{\alpha+1}\left(1 - \frac{1}{\nu+1}\right)^\alpha y^{\alpha+1}; \quad -1 < x \le y;\ y > 0;\ \nu > 0;$$

which implies

$$-(p-1)e_i(y_i - e_i)^{1-\frac{2}{p}} \ge -\frac{p}{2}\left(1 - \frac{p}{2(p-1)}\right)^{\frac{p-2}{p}} y_i^{\frac{2(p-1)}{p}}.$$

Combining the two inequalities we come to a lower bound for ith diagonal element of matrix D, and therefore

$$\frac{1}{2p}\frac{\partial^2 S}{\partial \beta^2} \ge T\sum y_i^{2-\frac{2}{p}} x_i x_i' - \left(2-\frac{2}{p}\right)\sum y_i^{1-\frac{2}{p}} e_i x_i x_i'.$$

where

$$T = T(p) = p\left(1 - 4^{\frac{1-p}{p}}\left(\frac{p-2}{p-1}\right)^{\frac{p-2}{p}}\right).$$

Hence, the LCL for the power regression in the case $p > 2$ is

$$S_{LC} = \frac{T^2 p^2}{4(p-1)^2}\,\lambda_{\min}\left\{\sum y_i^{2-\frac{2}{p}} x_i x_i' \,|\, \sum y_i^{1-\frac{2}{p}} x_i x_i';\ i = 1,\ldots,n\right\}.$$

Again, $\lambda_{\min}$ is calculated by the repeated eigenvalue problem considered in the previous section.

3. $p < 0$: Let us introduce the following function of x;

$$L(x) = (y-x)^{1-2/p}(-py + (1-2p)x); \quad x < y$$

where $y > 0$ is fixed. It is possible to show that for this function

$$L(x) > L(0) + x \cdot \left.\frac{dL}{dx}\right|_{x=0} = -py^{2-2/p} - 3(1-p)y^{1-2/p}x \tag{6.3}$$

for all x. Hence, applying the inequality (6.3) to $-D_{ii}(\theta)$ we obtain

$$-\frac{1}{2p}\frac{\partial^2 S}{\partial \theta^2} \cdot -p \sum y_i^{2-\frac{2}{p}} x_i x_i' - 3(1-p) \sum y_i^{1-\frac{2}{p}} e_i x_i x_i'.$$

Therefore, the local convexity level is

$$S_{LC} = \frac{p^2}{9(1-p)^2} \cdot \min \left\{ \sum y_i^{2-\frac{2}{p}} x_i x_i' \Big/ y_i^{1-\frac{2}{p}} x_i x_i';\ i = 1,\ldots,n \right\}. \tag{6.4}$$

Example. We consider an example from Bates and Watts (1988, p. 114) with Michaelis-Menten model (modified) $f(z;\mu) = \mu_1 z/(\mu_2 + z + \mu_3 z^2)$ which can be parameterized as power model with $p = -1$ as follows, $\theta_1 = 1/\mu_1; \theta_2 = \mu_2/\mu_1; \theta_3 = \mu_3/\mu_1$ and $x_1 = 1; x_2 = 1/z; x_3 = z$. The LSE is $(1.13 \cdot 10^{-5}; 2.08 \cdot 10^{-3}; 1.81 \cdot 10^{-7})$ with the minimum SS $1.38 \cdot 10^6$. The S_{LC} value for this problem, computed by formula (6.4), is $4.81 \cdot 10^6$.

7. Robust M-regression

It is well known that the LSE is not robust to outliers (Huber 1981). The method of robust M-regression is based on the minimization of sum of a function of residuals, different from quadratic,

$$W(\theta) = \sum \rho(y_i - \theta' x_i), \tag{7.1}$$

where $\rho = \rho(e)$ is a loss function with the properties: a) $\rho(0) = 0$; b) $\rho(e) = \rho(-e)$; c) $\rho'(e) \geq 0; e \geq 0$. As was pointed out by Huber (1981), when ρ is not convex, the minimization of $W(\theta)$ may cause problems because it may have several local minima: "Unless we are careful we may even get trapped in a local minimum of $W(\theta)$: The situation gets particularly acute in multiparameter regression". The reader is refered to Fig. 3 where a piece of surface (7.1) is displayed for a real life data (see Example below).

A number of authors addressed the question of optimal choice of the loss function and relevant comparisons have been made (e.g. Andrews et al. 1972). Particularly, it was suggested to choose ρ close to quadratic function near zero with asymptotes at $|e| = 1$: Probably, the simplest loss function of this kind is

$$\rho(e) = \frac{e^2}{c + e^2}; \tag{7.2}$$

where $c > 0$ is a given parameter which affects the sensitivity of the M-regression to outliers. Montgomery and Peck (1992), among others, consider other loss functions and discuss iterative algorithms for W minimization.

We start with finding regions where $W(\theta)$ is convex, then we construct sufficient global criteria. The Hessian matrix for function (7.1) is

$$\frac{\partial^2 W}{\partial \theta^2} = \sum \rho''(y_i - \theta' x_i) x_i x_i' \tag{7.3}$$

where in our case

$$\rho''(e) = \frac{2c(c - 3e^2)}{(c + e^2)^3}. \tag{7.4}$$

A simple global criterion for the M-regression can be constructed based on the GCGMI formulated in the previous section. Indeed, from (7.4) we notice that $\partial^2 W/\partial\theta^2$ is p.d. if $c - 3(y_i - a' x_i)^2 > 0$ for all $i = 1, \ldots, n$. Thus, on the convex set $L = \{\theta : |y_i - \theta' x_i| < \sqrt{c/3}; i = 1, \ldots, n\}$ the objective function W is convex. Now, to apply the GCCMI we need to find a lower bound for $W(\theta); \theta \in L$. In fact, if $\theta \in L$ then for some j we have $|y_j - \theta' x_j| \geq \sqrt{c/3}$ and consequently $W(\theta) \geq \rho(\sqrt{c/3}) = 1/4$. Therefore we infer that if θ_* is a local minimizer with $W(\theta_*) < 1/4$ then θ_* is the global minimizer of (7.1), i.e., the M-estimate.

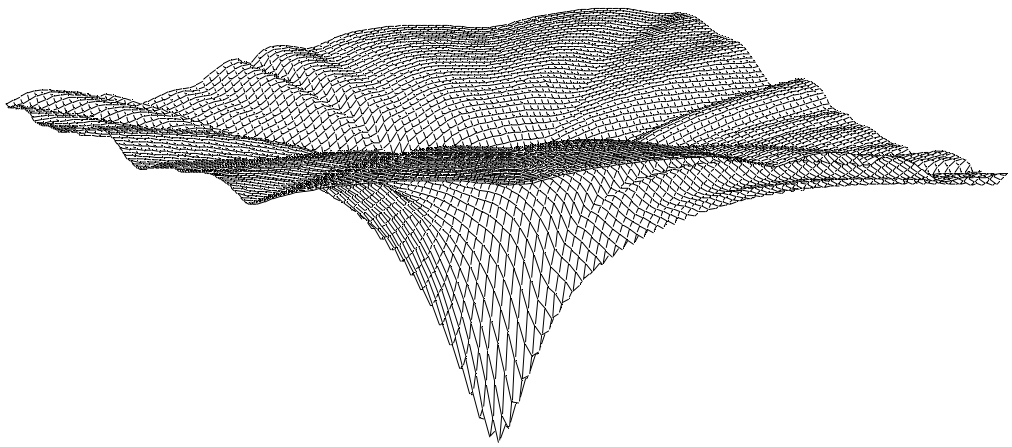

Fig. 3. A "mashroom" surface $W(\theta_1, \theta_2)$ for the M-regression in the neighborhood of a minimum (global?) for vending machine example.

Now we suggest a more powerful criterion. We start from seeking a lower linear bound for ρ''. It is easy to show that for function (7.2)

$$\rho''(0) = 2=c; \quad \rho''(\pm\sqrt{c=3}) = 0; \quad \rho''(e) \to 0 \text{ when } |e| \to 1:$$

The minimum of $\rho''(e)$ is attained at $\pm\sqrt{c}$ and equal $-1:5c$: We can find the minimal positive constant D such that $\rho''(e) \geq \rho''(0) - D|e|$ for all $e \in R^1$: Geometrically, for positive e; it corresponds to a line which goes through $(0; 2=c)$ and touches $\rho''(e)$ at some point. Clearly, this point is determined by the equation $-\rho'''(e) = (2=c - \rho''(e))=e$ which is equivalent to the cubic equation $u^3 + 4u^2 + 21u - 6 = 0$ where $u = e^2=c$: This equation has a unique real root u_* (approximately $u_* = 0:2708$): Hence $D = -\rho'''(\sqrt{u_* c}) = 3:492c^{-3=2}$; and the lower bound for ρ'' can be written as $\rho''(e) \geq 2=c - 3:492c^{-3=2}|e|$: Substituting it into (7.3) we come to a lower bound for the Hessian

$$\frac{\partial^2 W}{\partial \beta^2} \geq 3:492c^{-3=2} \left[0:5727\sqrt{c} \sum x_i x_i' - \sum |y_i - \beta' x_i| x_i x_i' \right]:$$

As follows from the previous section on the ellipsoid

$$E = \{\beta \in R^m : \sum (y_i - \beta' x_i)^2 < q\} \text{ where } q = (0:5727)^2 c \, \tau_{min} \tag{7.5}$$

and $\tau_{min} = \tau_{min}(\sum x_i x_i'; |x_i x_i'|; i = 1; \ldots; n)$; the function $W(\beta)$ is convex.

Now, in order to apply GCGMI, we need to find a lower bound for W outside ellipsoid (7.5). For this purpose we make use of the following simple optimization problem

$$\sum \min_{(v_i - \beta' x_i)^2 \geq q} |v_j - \beta' x_j| = \max(0; \gamma_j) \tag{7.6}$$

where $\gamma_j = \sqrt{(q - S_{min}) x_j (X'X)^{-1} x_j} - |v_j - a_{OLS}' x_j|$ and a_{OLS} is the OLS-estimate in regression $\{v_i\}$ on $\{x_i\}$; $q \geq S_{min} = \sum (v_i - a_{OLS}' x_i)^2$: Then if $\beta \in E$; as follows from (7.6). letting $v_i = y_i$

$$W(\beta) = \sum \frac{e_i^2}{c + e_i^2} \geq \sum \frac{\max^2(0; \gamma_i)}{c + \max^2(0; \gamma_i)} = W_g: \tag{7.7}$$

Based on this inequality the following sufficient global criterion can be formulated.

Criterion for the global minimum of $W(\beta)$. If for a local minimizer β_* we have $W(\beta_*) < W_g$ then β_* is the global minimizer, i.e., the true M-estimate.

Example. The above criterion is illustrated by a problem analyzing the vending machine service routes (Montgomery and Peck 1982, Example 9.4). The data consists of 25 observations where the dependent variable y is the delivery time (minutes), x_1 is the number of cases of product stocked and x_2 is the distance walked by the route driver (feet). To reduce

this problem to a two-parameter regression we eliminate the intercept term centering data: $\tilde{y}_i = y_i - \bar{y}$; $\tilde{x}_{i1} = x_{i1} - \bar{x}_1$; and $\tilde{x}_{i2} = x_{i2} - \bar{x}_2$: Then, the model is $\tilde{y}_i = \beta_1\tilde{x}_{i1} + \beta_2\tilde{x}_{i2} + \varepsilon_i$: We apply the loss function (7.2) with $c = 400$. The objective function, $W(\beta_1;\beta_2)$ is quite "wavy" (see Fig. 2), and possibly has several local minima. The minimization of (7.1) by iterative weighted least squares (Huber 1981) leads to a local minimum $\beta_{*1} = 1{:}594$; $\beta_{*2} = 0{:}014$ with $W(\beta_{*1};\beta_{*2}) = 0{:}544$: Is this the global minimum? To answer the question we apply the above criterion based on calculation (7.7). For our data $q = 795{:}2$ and $W_g = 1{:}13 > 0{:}544 = W(\beta_{*1};\beta_{*2})$: Therefore, the answer is positive { the found local minimum is the global one.

8. Exponential regression models

This type of nonlinear regression is popular, for instance, in econometrics (Kobb-Douglas production function, Judge et al. 1982) and chemistry (monomolecular growth curves, Seber Wild 1989). The regression function of the exponential model is de¯ned as $f_i(\beta) = \exp(\beta' x_i)$ where $\beta \in R^m$ and the system of vectors $\{x_1;\ldots;x_n\}$ has rank m: The Hessian of the SS for this regression is given by

$$\frac{\partial^2 S}{\partial \beta^2} = 2\sum (2e^{\beta' x_i} - y_i)x_i x_i': \tag{8.1}$$

We assume here $y_i > 0$ for all $i = 1;\ldots;n$: As follows from (8.1), on the polyhedra

$$P = \{\beta \in R^m : \beta' x_i > \ln y_i - \ln 2; \quad i = 1;\ldots;n\} \tag{8.2}$$

the SS is convex. Now we apply GCGMI to construct a simple global criterion. For this we have to ¯nd a lower bound for the SS outside P: In fact, if $\beta \notin P$ then for some j we have $y_j - e^{\beta' x_j} \geq 0{:}5y_j > 0$ which implies

$$S(\beta) = \sum (y_i - e^{\beta' x_i})^2 \geq (y_j - e^{\beta' x_j})^2 \geq \frac{1}{4}y_j^2 \geq \frac{1}{4}\min_i\{y_i^2\}:$$

Therefore, we can construct the following quick global criterion: if β_* is a local minimizer of the SS and $S(\beta_*) < :25\min\{y_i^2\}$; then β_* is the LSE.

Now we ¯nd a more precise LCL. Let us denote $e_i = y_i - e^{\beta' x_i}$; then $2e^{\beta' x_i} - y_i = y_i^2 - 3e_i y_i + 2e_i^2$; and

$$\frac{\partial^2 S}{2\partial \beta^2} = \sum y_i^2 x_i x_i' - \sum e_i y_i x_i x_i' - 2\sum e_i(y_i - e_i)x_i x_i': \tag{8.3}$$

For the third term in (8.3) we apply an elementary inequality $e_i(y_i - e_i) \leq y_i^2/4$ which leads us to the following lower bound for the half Hessian

$$\frac{\partial^2 S}{2\partial \beta^2} \geq \frac{1}{2}\sum y_i^2 x_i x_i' - \sum e_i y_i x_i x_i':$$

Therefore the LCL can be de¯ned as

$$S_{LC} = \frac{1}{2}\,'_{min}\Big(\sum y_i^2 x_i x_i^{\prime} \;|\; y_i x_i x_i^{\prime};\ i = 1;:::;n\Big): \tag{8.4}$$

Hence, for any parameter with the SS less than S_{LC} de¯ned by (8.4) the Hessian (8.1) is positive de¯nite.

Now we construct a global criterion. Since function e^v is convex $e^v \geq e^{v_0}(v - v_0) + e^{v_0}$ for any v and v_0: Applying this inequality to (8.1) with $v = \theta^{\prime}x_i$ and $v_0 = \ln y_i$ we obtain

$$\frac{\partial^2 S}{2\partial\theta^2} \geq \sum y_i x_i x_i^{\prime} - 2\sum (\ln y_i - \theta^{\prime}x_i) y_i x_i x_i^{\prime}:$$

Therefore on the ellipsoid

$$E = \{\theta \in R^m : \sum (\ln y_i - \theta^{\prime}x_i) < q\} \text{ where } q = \frac{1}{4}\,'_{min}\Big(\sum y_i x_i x_i^{\prime} \;|\; y_i x_i x_i^{\prime};\ i = 1;:::;n\Big)$$

the SS is convex. In order to apply GCGMI we need to ¯nd a lower bound for $S(\theta)$ when θ lies outside E: For this purpose we again use the inequality (7.6), letting $v_i = \ln y_i$: Since $|\ln y_j - \theta^{\prime}x_j| \leq \varepsilon_j$ for $\varepsilon_j \geq 0$ implies $(y_j - e^{\theta^{\prime}x_j})^2 \geq y_j^2(1 - e^{\pm\varepsilon_j})^2$ we obtain

$$S(\theta) \geq \sum_{\varepsilon_j > 0} y_j^2(1 - e^{\pm\varepsilon_j})^2 = S_g; \qquad \theta \in E: \tag{8.5}$$

Hence, if θ_* is a local minimizer in exponential model with the SS less than S_g then θ_* is the global minimizer.

Example. We borrow data from Bates and Watts (1989, p. 281) on tetracycline concentration (¹g/ml) versus time (hr). Thus, if y denotes tetracycline concentration and t time we suggest the following model

$$y_i = \exp(\theta_1 + \theta_2 t_i^{\pm 1} + \theta_3 t_i) + \varepsilon_i; \qquad i = 1;:::;n = 9:$$

The nonlinear estimation starting from the OLS-estimates in the linear regression $\ln y_i$ on $\theta_1 + \theta_2 t_i^{\pm 1} + \theta_3 t_i$ comes to a local minimizer $\hat{\theta} = (1.33; -0.163; -1.56)$ with $SS_{min} = 0.0142$. Is $\hat{\theta}$ the true LSE, i.e., did we ¯nd the global minimum of the SS? The $'-$value is 0.355 and $S_g = 0.476$. Since $SS_{min} < S_g$ we infer that $\hat{\theta}$ is the true least-squares estimate.

9. Conclusions

The existing algorithms for sum of squares minimization, such as Gauss-Newton or Marquardt, can ¯nd only a local minimum. Unlike linear regression the SS for nonlinear regression may have several local minima. Moreover, if a nonlinear model cannot be reparameterized to linear, the probability to have at least two local minima is positive. Therefore, almost

in all cases of nonlinear estimation there is a doubt that the model was estimated correctly, i.e., the found local minimizer is in fact the Least Squares Estimate. The current paper aims to provide some guidelines in constructing criteria which verify that the local minimum is the global one. Notice, these criteria are sufficient, not necessary. In other words if criteria fail it does not mean the SS has two local minima and the found minimum is not the global one. Our strategy is to express the properties of the SS in terms of descended levels:

1. Existence Level. This level was introduced in Demidenko (1981) and recently illustrated through different types of nonlinear models in Demidenko (1996). Hence, if $\overline{S}_E$ is an EL and θ_0 is an initial parameter vector such that $S(\theta_0) < \overline{S}_E$ then the LSE exists.

2. Local Convexity Level. This level is equal to the minimum of squared radius of the full curvature of the expectation surface. Hence, if S_{LC} is a level of LC and $S(\theta) < S_{LC}$ then it is guaranteed that the Hessian of SS is positive definite at θ.

3. Global Minimum Level. If S_G is such a level and θ_* is a local minimizer such that $S(\theta_*) < S_G$ then θ_* is the global minimizer, i.e., the LSE.

One need to understand that except rare cases we are not able to find the (exact) upper local convexity level because it would lead to another minimization problem, perhaps even more severe than the SS minimization itself. However, we succeed to find a lower bound. Much creativity should apply to find a satisfactory lower bound and examples considered in this paper hopefully illustrate this. However, much work should be done to develop these techniques for other, more complicated nonlinear regression models. Global criteria must be a part of software on nonlinear estimation.

ACKNOWLEDGMENT

This work was supported by National Cancer Institute grants CA52192 and CA61108.

10. Appendix. Proofs

The following auxiliary result will be used, Demidenko (p. 85, 1989).

Lemma 10.1. A multivariate regression is intrinsically nonlinear if and only if there exists θ_0 such that the system of $m(m+3)/2$ vectors from R^n

$$\left(\frac{\partial f(\theta_0)}{\partial \theta_j}; \frac{\partial^2 f(\theta_0)}{\partial \theta_j \theta_k}; \quad j = 1, \ldots, m; k \le j \right) \tag{10.1}$$

has full rank.

Also, the following fundamental result of optimization theory will be used.

Lemma 10.2. Let $F(u)$ be a twice differentiable function of $u \in R^m$ such that $\|u\| \to \infty$ implies $F(u) \to \infty$. If there exists u_0 such that $\partial F(u_0)/\partial u = 0$ and $\partial^2 F(u_0)/\partial u^2$ is a negative definite matrix then $F(u)$ has at least two local minima on R^m.

As pointed out by Mäkeläinen, Schmidt and Styan (1981) this result is a straightforward consequence of Morse theory, however they provide an elementary proof for a similar result. A more general results is proved in Demidenko (1989).

Proof of Theorem 2.1. The idea of the proof is to use the implicit function theorem by showing that the Jacobian matrix for an introduced mapping has full rank. It provides that, under conditions of the Theorem, observations y with multimodal SS fill an open set in R^n. Clearly, that implies the probability to have at least two local minima of SS is positive, provided that the density of the error distribution is positive.

Now we realize this plan in some detail. Let us denote the $n \times m_*$ matrix with the columns (10.1) as $M(\theta_0)$ where $m_* = (m(m+3)/2)$. As follows from Lemma 10.1 it has full rank. Without loss of generality one can assume that the upper $m_* \times m_*$ part of this matrix is nonsingular. Since the elements of matrix M are continuous functions of θ the upper matrix has full rank in some neighborhood $V = V(\theta_0)$. Further, let us part matrix $M(\theta)$ as follows

$$M(\theta) = M = \begin{bmatrix} G_1 & U_1 \\ G_2 & U_2 \end{bmatrix};$$

where G_1 is the $m_* \times m$ matrix. Then $G = [G_1', G_2']'$ is the matrix of first derivatives and $U = [U_1', U_2']'$ is the matrix of second derivatives (argument θ is omitted). As was noticed above, $\det[G_1; U_1] \neq 0$. Further, let us consider the set of all symmetric positive definite $m \times m$ matrices M_+. Clearly, any matrix D from M_+ is determined by $m(m+1)/2$ numbers. As follows from Lemma 10.2, the SS has at least two local minima if for some θ we have $\partial S(\theta)/\partial\theta = 0$ and $\partial^2 S(\theta)/\partial\theta^2 = -D$. We aim to show that when θ takes values from V and D is p.d. the solutions of the two last equations, y fill a certain $n-$dimentional region in R^n. Since the Hessian is a symmetric matrix we can apply the vech function (e.g., Magnus and Neudecker 1988) to $\partial^2 S(\theta)/\partial\theta^2 = -D$ and therefore come to an equivalent vector equation

$$-\sum \operatorname{vech}\left(\frac{\partial^2 f_i}{\partial\theta^2}\right)(y_i - f_i(\theta)) = -r - d$$

where $r = \operatorname{vech}(G'G)$ and $d = \operatorname{vech}(D)$. Thus, $\partial S(\theta)/\partial\theta = 0$ and $\partial^2 S(\theta)/\partial\theta^2 = -D$ are equivalent to the following system of linear equations for y:

$$G_1'(y_1 - f_1) + G_2'(y_2 - f_2) = 0; \tag{10.2}$$

$$U_1(y_1 - f_1) + U_2(y_2 - f_2) = r + d; \quad \theta \in V; D \in M \tag{10.3}$$

where $y_1 = (y_1, ..., y_{m_*})'$; $y_2 = (y_{m_*+1}, ..., y_n)'$ and the similar partition for f: Now we construct the following mapping:

$$a : (\theta', d', y_2') \to (y_1', y_2'), \qquad \theta \in V,\ D \in M_+,\ y_2 \in R^{n-m_*} \tag{10.4}$$

where y_1 is the solution to (10.2,10.3) given θ, d, y_2. This solution always exists because matrix $[G_1; U_1]$ has full rank. Our current aim is to find the Jacobian of y_1 as a function of θ; d and show it is nonsingular. Differentiating (10.2) with respect to θ and d we obtain

$$G_1' \frac{\partial y_1}{\partial \theta} + D = 0, \qquad G_1' \frac{\partial y_1}{\partial d} = 0.$$

Differentiating (10.3) with respect to θ and d we obtain

$$U_1' \frac{\partial y_1}{\partial \theta} + R = \frac{\partial r}{\partial \theta}, \qquad U_1' \frac{\partial y_1}{\partial d} = I, \quad \text{where } R = \frac{\partial U_1'}{\partial \theta} y_1 + \frac{\partial U_2 (y_2 - f_2)}{\partial \theta}.$$

Combining these equations we obtain

$$\det \begin{bmatrix} \frac{\partial y_1}{\partial \theta}, \frac{\partial y_1}{\partial d} \end{bmatrix} = \det ([G_1; U_1]')^{-1} \cdot \det \begin{bmatrix} -D & 0 \\ \frac{\partial r}{\partial \theta} - R & I \end{bmatrix} = -\det ([G_1; U_1]')^{-1} \det(D) \neq 0,$$

which proves the Jacobian is nonsingular. Now it suffices to apply the implicit function theorem for mapping (10.4). It has a nonsingular Jacobian and, consequently, in some neighborhood has inverse. Therefore, the 'observations', $y = (y_1; y_2)$ fill an $n-$dimentional region and for each y from it the SS has at least two local minima. Since the density of the distribution of the error term is positive, the probability to fall in this region is positive as well.

References

[1] Abramowitz, M., and Stegun, I.A. (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards. New York.

[2] Amari, S.-I. (1985). Differential Geometrical Methods in Statistics. Lecture Notes in Statistics, 28, Springer, Berlin.

[3] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). Robust Estimates of Location. Princeton University Press, Princeton. New Jersey.

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

[5] Chong, E.K., and Zak, S.H. (1996). An Introduction to Optimization. Wiley. New York.

[6] Demidenko, E. (1981). Linear and Nonlinear Regression (in Russian). Nauka. Moscow.

[7] Demidenko, E. (1989). Optimization and Regression (in Russian). Nauka. Moscow.

[8] Demidenko, E. (1996).On the existence of the least squares estimate in nonlinear growth curve models of exponential type. Communications in Statistics, Theory and Methods, 25, 159-182.

[9] Ezekiel, M., and Fox, K.A. (1959). Methods of correlation and regression analysis, linear and curvilinear. Wiley. New York. 1959.

[10] Floudas, C.A. and Pardalos, P.M. (1992). Recent Advances in Global Optimization. Princeton University Press. New Jersey.

[11] Gallant, A.R. (1987). Nonlinear Statistical Methods. Wiley. New York.

[12] Hansen, E. (1992). Global Optimization Using Interval Analysis. Dekker. New York.

[13] Horst, R. and Pardalos, P.M. (1995). Handbook of Global Optimization. Kluwer. Dordrecht.

[14] Horst, R. and Tuy, H. (1996). Global Optimization. Deterministic Approaches. Springer. New York.

[15] Huber, P.J. (1981). Robust Statistics. Wiley. New York.

[16] Judge, G.C., Hill, R.C., Gri±ths, W.E., Lütkepohl, H., and Lee, T.-C.(1982). Introduction to the Theory and Practice of Econometrics. Wiley. New York.

[17] Kearfort, R.B. (1996). Rigorous Global Search. Kluwer. Dordrecht.

[18] MäkeläInen, T., Schmidt, K., and Styan, G.P.H. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in ¯xed-size samples. Annals of Statistics, 9, 758-767.

[19] Mockus, J. (1989). Bayesian Approach to Global Optimization.. Theory and Application. Kluwer. Dordrecht.

[20] Montgomery, D.C., and Peck, E.A. (1992). Introduction to Linear Regression Analysis. Wiley. New York.

[21] Nakamura, T. (1984). Existence theorems of a maximum likelihood estimate from a generalized censored data sample. Annals of the Institute of Statistical Mathematics, 36, 375-393.

[22] Ortega, J.M., and Rheinboldt, W.C. (1970). Iterative Solution of Nonlinear Equations in Several Variables. Academic Press. New York.

[23] Pazman, A. (1984). Nonlinear least squares - uniqueness versus ambiguity. Math. Operationsforschung Stat., ser. Statistics, 15, 323-336.

[24] Pazman, A. (1993). Nonlinear Statistical Models, Kluwer. Dordrecht.

[25] Pinter, J.D. (1996). Global Optimization in Action. Kluwer. Dordrecht.

[26] Rockafellar, R.T. (1970). Convex Analysis, Princeton University Press. New Jersey.

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

[28] Webster, R. (1994). Convexity, Oxford University Press. Oxford.

[29] Zhigljavsky, A.A. (1991). Theory of Global Random Search. Kluwer. Dordrecht.