Preponderantly increasing/decreasing data in regression analysis

Darija Marković[1,*]

[1] *Department of Mathematics, J. J. Strossmayer University of Osijek,*
Trg Ljudevita Gaja 6, 31000 Osijek, Croatia
E-mail: ⟨*darija@mathos.hr*⟩

Abstract. For the given data (w_i, x_i, y_i), $i = 1, \dots, n$, and the given model function $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters, the goal of regression analysis is to obtain estimator $\boldsymbol{\theta}^*$ of the unknown parameters $\boldsymbol{\theta}$ such that the vector of residuals is minimized in some sense. The common approach to this problem of minimization is the least-squares method, that is minimizing the L_2 norm of the vector of residuals. For nonlinear model functions, what is necessary is finding at least the sufficient conditions on the data that will guarantee the existence of the best least-squares estimator. In this paper we will describe and examine in detail the property of preponderant increase/decrease of the data, which ensures the existence of the best estimator for certain important nonlinear model functions.

Keywords: regression analysis, nonlinear least squares, existence problem, preponderant increase/decrease property, Chebyshev inequality

Received: October 3, 2016; accepted: December 11, 2016; available online: December 30, 2016

DOI: 10.17535/crorr.2016.0018

1. Introduction

Regression analysis is one of the most important tools for data analysis. For the given model function $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters, the regression model can be written in the form

$$y_i = f(x_i; \boldsymbol{\theta}) + \varepsilon_i, \qquad i = 1, \dots, n.$$

Here x_i denote the values of the independent variable, y_i denotes the value of the dependent variable and ε_i are random errors. The goal of regression analysis is to obtain the estimator $\boldsymbol{\theta}^*$ of the unknown parameters $\boldsymbol{\theta}$ base on the given experimental or empirical data (w_i, x_i, y_i), where $w_i > 0$ are in advance given data weights. Inverse-variance weighting is typically used in statistical literature, but various weighting methods are used to calculate weights (see [12]). The errors ε_i are usually assumed to be normally distributed with mean zero and constant variance.

[*] Corresponding author.

In the most general way, we can divide regression models into two classes: linear and nonlinear regression models. Linear regression models are those that are linear in all parameters and are often used in applied sciences due to their simplicity. However such models are usually nonrealistic. Nonlinear regression models usually appear when there is some physical explanation of the relationship between the dependent and the independent variable in a particular functional form. Practical introductions to nonlinear regression including numerous data examples are given by Ratkowsky in [11] and by Bates and Watts in [1]. A more extensive treatment of nonlinear regression methodology is given by Seber and Wild in [15].

The vector of unknown parameters $\boldsymbol{\theta}$ in the regression model should be estimated from the given data by minimizing a suitable goodness-of-fit expression with respect to $\boldsymbol{\theta}$. The most popular criterion in applied sciences is based on minimization of the weighted sum of squared residuals, that is by minimizing the following expression

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i (y_i - f(x;\boldsymbol{\theta}))^2 = \sum_{i=1}^{n} w_i \varepsilon_i^2.$$

This approach is known as weighted least squares (LS) method.

It is well known that for a linear least squares problem a closed form solution exists and can be easily obtained. Numerical methods for solving the nonlinear LS problem are described in [1, 11, 12, 15]. Before performing iterative minimization of the sum of squares, the questions remains as to whether the least squares estimate (LSE) exists. For the case of nonlinear LS problems, answering this question is exceptionally difficult (see [1, 11, 12, 15]). Therefore, in order to guarantee the existence of a minimum of the functional S (i.e. the existence of the LSE) it is necessary to require that the data (w_i, x_i, y_i), $i = 1, \dots, n$, satisfy some additional conditions. It has been shown that property of preponderant increase/decrease, which will be described in more details in the next section, has an important role in analyzing the existence of the LSE (see e.g. [2, 3, 5, 6, 7, 13, 14]). Our main theoretical result is the useful Theorem 1 that describes increasing/decreasing data by preponderantly increasing/decreasing data. Its application in some nonlinear least squares existence problems is illustrated in Section 3.

2. Property of preponderant increase/decrease of the data

Let us first recall some important and useful definitions and results. Throughout the entire paper we will suppose that we are given data (x_i, y_i), $i = 1, \dots, n$, such that

$$x_1 \le x_2 \le \cdots \le x_n \quad \text{and} \quad x_1 < x_n. \tag{1}$$

We will say that the data (x_i, y_i) are *increasing* if

$$y_1 \le y_2 \le \cdots \le y_n \quad \text{and} \quad y_1 < y_n.$$

Similarly, we will say that the data are *decreasing* if

$$y_1 \ge y_2 \ge \cdots \ge y_n \quad \text{and} \quad y_1 > y_n.$$

If $y_1 = y_2 = \cdots = y_n$ we say that the data are *constant* or *stationary*.

The property of preponderant increase/decrease can be described by the Chebyshev inequality. This inequality is usually stated as follows (see [4, 10]): Let $a_1 \leq a_2 \leq \cdots \leq a_n$ and $b_1 \leq b_2 \leq \cdots \leq b_n$. Then

$$n \sum_{i=1}^{n} a_i b_i \geq \sum_{i=1}^{n} a_i \sum_{i} b_i.$$

The above Chebyshev inequality can be extended to include positive weights p_i,

$$\sum_{i=1}^{n} p_i \sum_{i=1}^{n} p_i a_i b_i \geq \sum_{i=1}^{n} p_i a_i \sum_{i=1}^{n} p_i b_i. \tag{2}$$

The Chebyshev inequality can be reversed: If $a_1 \leq a_2 \leq \cdots \leq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$ then

$$\sum_{i=1}^{n} p_i \sum_{i=1}^{n} p_i a_i b_i \leq \sum_{i=1}^{n} p_i a_i \sum_{i=1}^{n} p_i b_i. \tag{3}$$

If at least two of the a_i and at least two of the b_i are distinct then both inequalities (2) and (3) are strict.

For any two finite sequences of real numbers $\boldsymbol{a} = (a_1, \ldots, a_n)$ and $\boldsymbol{b} = (b_1, \ldots, b_n)$ and any strictly positive and finite sequence of real numbers $\boldsymbol{p} = (p_1, \ldots, p_n)$, the so-called Korkine's identity holds (see [9]):

$$\sum_{i=1}^{n} p_i \sum_{i=1}^{n} p_i a_i b_i - \sum_{i=1}^{n} p_i a_i \sum_{i=1}^{n} p_i b_i = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j (a_i - a_j)(b_i - b_j). \tag{4}$$

This equality is easy to check directly by rewriting the right-hand side of the equation.

Before stating the definition of preponderantly increasing/decreasing data, let us first recall that, for the given data (w_i, x_i, y_i), $i = 1, \ldots, n$, the corresponding weighted regression line is given by

$$y = a^* x + b^*,$$

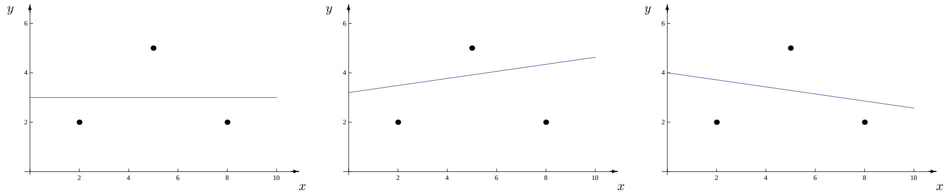
where

$$\begin{aligned} a^* &= \frac{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2}, \\ b^* &= \frac{\sum_{i=1}^{n} w_i y_i - a^* \sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}. \end{aligned}$$

Definition 1 (see [13]). *The data (w_i, x_i, y_i), $i = 1, \ldots, n$, are said to have the preponderant increase (respectively decrease) property if the slope a^* of the associated weighted linear trend is positive (respectively negative). If the slope is equal to zero, then the data is said to be preponderantly stationary.*

The property of preponderant increase/decrease is also referred to as an essential increase/decrease property (see e.g. [2]). In general, this property depends on both data (x_i, y_i) and weights w_i. We will illustrate this in the next simple example.

Example 1. *Let the data* $\boldsymbol{x} = (2,5,8)$ *and* $\boldsymbol{y} = (2,5,2)$ *be given. We will make the following choices of weight:* $\boldsymbol{w}_1 = (1,1,1)$, $\boldsymbol{w}_2 = (1,1,\frac{1}{2})$ *and* $\boldsymbol{w}_3 = (\frac{1}{2},1,1)$. *For the first choice of weights, the slope of the regression line is equal to zero, hence the data are preponderantly stationary. For the second choice of weights, the slope of the regression line is equal to* $\frac{1}{7}$, *and the data have the property of preponderant increase. The last choice of weights for the slope gives the value of* $-\frac{1}{7}$, *and hence in this case the data have the property of preponderant decrease.*



(a) $(\boldsymbol{w}_1, \boldsymbol{x}, \boldsymbol{y})$, $y = 3$ (b) $(\boldsymbol{w}_2, \boldsymbol{x}, \boldsymbol{y})$, $y = \frac{1}{7}x + \frac{16}{5}$ (c) $(\boldsymbol{w}_3, \boldsymbol{x}, \boldsymbol{y})$, $y = -\frac{1}{7}x + 4$

Figure 1: *The data and associated linear trends for different choices of weights*

Given that the denominator of a^* is strictly positive as x_i satisfy (1), the data will be preponderantly increasing if and only if the following condition is satisfied

$$Q(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i > 0.$$

By using (4), that condition can be written in the following form:

$$Q(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j (x_i - x_j)(y_i - y_j) > 0. \tag{5}$$

The above discussion is stated in the following proposition (see also [13]).

Proposition 1. *The data* (w_i, x_i, y_i), $i = 1, \ldots, n$, *have the property of preponderant increase if and only if the Chebyshev inequality holds:*

$$\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i > \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i.$$

Similarly, property of preponderant decrease is equivalent to the opposite inequality.

The following results are direct consequences of Proposition 1 (see also [2]).

Proposition 2. *Effect of some transformation of the data.*

(a) *If the data* (w_i, x_i, y_i)*,* $i = 1, \dots, n$*, are preponderantly increasing (decreasing), then the data* $(w_i, -x_i, y_i)$ *are preponderantly decreasing (increasing).*

(b) *If the data* (w_i, x_i, y_i)*,* $i = 1, \dots, n$*, are preponderantly increasing (decreasing), then the data* $(w_i, x_i, -y_i)$ *are preponderantly decreasing (increasing).*

(c) *If the data* (w_i, x_i, y_i)*,* $i = 1, \dots, n$*, are preponderantly increasing (decreasing) and* $0 < x_1$*, then the data* $(w_i x_i, \tilde{x}_i, y_i)$*, where* $\tilde{x}_i = \frac{1}{x_i}$*, are preponderantly decreasing (increasing).*

(d) *If the data* (w_i, x_i, y_i)*,* $i = 1, \dots, n$*, are preponderantly increasing (decreasing) and* $0 < y_1$ $(0 < y_n)$*, then the data* $(w_i y_i, x_i, \tilde{y}_i)$*, where* $\tilde{y}_i = \frac{1}{y_i}$*, are preponderantly decreasing (increasing).*

In regression analysis the most often case is when values of independent variable are distinct, i.e. $x_1 < x_2 < \cdots < x_n$. The following theorem treats just that case.

Theorem 1. *Suppose we are given the non-constant data* (x_i, y_i)*,* $i = 1, \dots, n$*, such that* $x_1 < x_2 < \cdots < x_n$*. The data are increasing (or decreasing) if and only if the data* (w_i, x_i, y_i) *have the property of preponderant increase (or preponderant decrease) for any choice of weights* $w_i > 0$*,* $i = 1, \dots, n$*.*

Proof. We will prove the Theorem 1 only for the case when data are increasing; the proof for the case when data are decreasing would be done in a similar way. Therefore suppose that the data (x_i, y_i), $i = 1, \dots, n$, are increasing. Then

$$(x_i - x_j)(y_i - y_j) \geq 0.$$

Since the data are increasing, strict inequality appears for at least one pair (i, j). Due to this fact, for any choice of weights $w_i > 0$ inequality (5) holds, and by definition, the data have the property of preponderant increase.

Suppose that the data (w_i, x_i, y_i), $i = 1, \dots, n$ are preponderantly increasing for any choice of weights $w_i > 0$, $i = 1, \dots, n$. Let $i_1, i_2 \in \{1, \dots, n\}$ such that $i_1 < i_2$. It is necessary to show that $y_{i_1} \leq y_{i_2}$ and $y_1 < y_n$. To do this, for each $k \in \mathbb{N}$ let us define

$$w_i := \begin{cases} \frac{1}{k}, & i \in \{1, \dots, n\} \setminus \{i_1, i_2\} \\ 1, & i \in \{i_1, i_2\}. \end{cases}$$

By assumption, we have

$$\begin{aligned} Q(\boldsymbol{w},\boldsymbol{x},\boldsymbol{y}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j (x_i - x_j)(y_i - y_j) \\ &= \frac{1}{k^2} \sum_{\substack{i=1\\ i\neq i_1,i_2}}^{n} \sum_{\substack{j=1\\ j\neq i_1,i_2}}^{n} (x_i - x_j)(y_i - y_j) \\ &\quad + \frac{1}{k} \sum_{\substack{j=1\\ j\neq i_2}}^{n} (x_{i_1} - x_j)(y_{i_1} - y_j) \\ &\quad + \frac{1}{k} \sum_{\substack{j=1\\ j\neq i_1}}^{n} (x_{i_2} - x_j)(y_{i_2} - y_j) \\ &\quad + 2(x_{i_2} - x_{i_1})(y_{i_2} - y_{i_1}) \\ &> 0, \qquad \forall k \in \mathbb{N}, \end{aligned}$$

wherefrom by passing to the limit when $k \to \infty$ we obtain

$$(x_{i_2} - x_{i_1})(y_{i_2} - y_{i_1}) \geq 0$$

Since $x_{i_2} - x_{i_1} > 0$, we have $y_{i_2} - y_{i_1} \geq 0$. It remains to show that $y_1 < y_n$. Indeed, if $y_1 = y_n$, then $y_1 = y_2 = \cdots = y_n$, implying that the slope of the weighted linear trend is equal to zero which contradicts the assumption. □

Remark 1. *The following result which was used in [3] for proof of the existence of the LSE can now be viewed as a consequence of Theorem 1 for a special choice of weights:*
Suppose that the data (w_i, x_i, y_i), $i = 1, \ldots, n$, are such that $0 < x_1 < x_2 < \cdots < x_n$ and $w_i > 0$, $i = 1, \ldots, n$. Then

(i) If the sequence $(y_1, \ldots, y_n)$ increases, then

$$\sum_{i=1}^{n} \frac{w_i}{x_i} \sum_{i=1}^{n} w_i y_i - \sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i \frac{y_i}{x_i} \geq 0, \tag{6}$$

with the inequality holding for $y_1 < y_n$.

(ii) If the sequence $(y_1/x_1, \ldots, y_n/x_n)$ decreases, then

$$\sum_{i=1}^{n} w_i y_i x_i \sum_{i=1}^{n} w_i x_i^3 - \sum_{i=1}^{n} w_i x_i^2 \sum_{i=1}^{n} w_i y_i x_i^2 \geq 0, \tag{7}$$

with the inequality holding for $y_1/x_1 > y_n/x_n$.

Condition (6) means that the data $(w_i/x_i, x_i, y_i)$, $i = 1, \ldots, n$, have the property of preponderant increase, and similarly, condition (7) means that the data $(w_i x_i^2, x_i, y_i/x_i)$, $i = 1, \ldots, n$, have the property of preponderant decrease.

3. Some applications of preponderant increase property in regression analysis

In this section we will give a brief overview of a few important nonlinear least square regression models. It has been shown that for these models the property of preponderant increase/ decrease is crucial for the existence of the LSE.

Example 2. *The mathematical model described by an exponential function*

$$f(x;b,c) = be^{cx}, \quad b,c \in \mathbb{R}$$

or a linear combination of such functions is often used in applied research, e.g. biology, chemistry, electrical engineering, economy, astronomy, nuclear physics, etc. In [7], it was shown that the LSE exists, provided the data satisfy either the condition of preponderant increase or the condition of preponderant decrease.

Example 3. *The generalized logistic function (or asymmetric S function)*

$$f(x;b,c) = \frac{A}{(1+be^{-c\gamma x})^{1/\gamma}}, \quad A,b,c,\gamma > 0$$

occurs frequently in various applied areas, such as biology, marketing, economics, etc. The existence of the LSE for the generalized logistic function is considered in [8], where it was proved that the LSE exists if in addition to the natural condition on the data, it is enough to require that the data are preponderantly increasing.

Example 4. *The Michaelis-Menten enzyme kinetic model*

$$f(x;a,b) = \frac{ax}{b+x}, \quad a,b > 0,$$

is widely used in biochemistry, pharmacology, biology and medical research. The following theorem gives sufficient conditions for the existence of the LSE; the proof can be found in [3].

Theorem 2. *Let the data (w_i, x_i, y_i), $i = 1,\dots,m$, $m \geq 3$, be given, such that $0 < x_1 \leq x_2 \leq \ldots \leq x_m$, $x_1 < x_m$ and $y_i > 0$, $i = 1,\dots,m$. If the data fulfill inequalities (6) and (7), then the LS estimate for the Michaelis-Menten function exists.*

4. Conclusion

The aim of this paper was to thoroughly investigate the property of preponderant increase/decrease of data. This property was found very useful in solving the existence problem for some important nonlinear model functions. The theoretical improvement of results given in [13] is stated in Theorem 1; that is the description of the increasing/decreasing data by preponderantly increasing/decreasing data.

Acknowledgement

I am grateful to the anonymous reviewers and the editor for their valuable suggestions, constructive comments and their efforts in improving this paper.

References

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

[2] Dubeau, F. and Mir, Y. (2007). On discrete least squares polynomial fit, linear spaces and data classification. J. Math. Stat., 3, 222-227.

[3] Hadeler, K.P., Jukić, D. and Sabo, K. (2007). Least squares problems for Michaelis Menten kinetics. Math. Meth. Appl. Sci., 30, 1231-1241.

[4] Hardy, G.H., Littlewood, J.E. and Pòlya, G. (1988). Inequalities, 2nd ed. Cambridge: Cambridge University Press.

[5] Jukić, D., Sabo, K. and Scitovski, R. (2007). Total least squares fitting Michaelis-Menten enzyme kinetic model function. J. Comput. Appl. Math., 201, 230-246.

[6] Jukić, D. and Scitovski, R. (1998). Existence results for special nonlinear total least squares problem. J. Math. Anal. Appl., 226, 348-363.

[7] Jukić, D. and Scitovski, R. (1997). Existence of optimal solution for exponential model by least squares. J. Comput. Appl. Math., 78, 317-328.

[8] Jukić, D. and Scitovski, R. (1996). The existence of optimal parameters of the generalized logistic function. Appl. Math. Comput., 77, 281-294.

[9] Korkine, A.N. (1882). Ob' odnom opredelennom integrale, (Russian). Math. Sbornik, 10, 571-572.

[10] Mitrinović, D.S., Pečarić, J. and Fink, A.M. (1993). Classical and New Inequalities in Analysis. Dordrecht: Kluwer.

[11] Ratkowsky, D.A. (1989). Handbook of Nonlinear Regression Models. New York: Dekker.

[12] Ross, G.J.S. (1990). Nonlinear Estimation. New York: Springer Verlag.

[13] Scitovski, R. (1988). Condition of preponderant increase and Tchebycheffs inequality. In Jovanović, B. S. (Ed.). Proceedings of VI. Conference on Applied Mathematics (pp. 189-194). Belgrade.

[14] Scitovski, R. (1988). Some special nonlinear least squares problems. Radovi matematički, 4, 279-298.

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