COMPUTATIONAL MATHEMATICS

MSC 62J05 **DOI: 10.14529/jcem180302**

ANALYSIS OF ALGORITHMS FOR STABLE ESTIMATION OF COEFFICIENTS OF MULTIPLE LINEAR REGRESSION MODELS

A. A. Azaryan, Yerevan State University, Yerevan, Republic of Armenia, a.a.azaryan@gmail.com

Computational experiments on model data were performed in order to study the effectiveness of the algorithms for realization of the least absolute deviations (LAD) method and the generalized method of the least absolute deviations (GLAD) when estimating the parameters of multiple linear regression models based on descent through the nodal straight lines. In addition, a comparative analysis of the algorithms of descent through nodal straight lines for LAD and GLAD with known exact and approximate methods to solve tasks (2) and (3) was carried out.

Keywords: linear regression model; the least absolute deviations method; generalized; computational complexity; comparative analysis.

Introduction

One of the most common tasks in the statistical processing of experimental findings is to estimate the unknown coefficients of multiple linear regression model [1]:

$$y_i = a_1 + a_2 x_{i2} + a_3 x_{i3} + \cdots + a_m x_{im} + \varepsilon_i, \quad i = \overline{1,n}\ , \tag{1}$$

where $y_i, i = \overline{1,n}$ are observed values of the dependent variable; $a_j, j = \overline{1,m}$ are unknown coefficients of multiple linear regression; $x_{ij}, j = \overline{1,m}, i = \overline{1,n}$ are values of explanatory (independent) variables; $\varepsilon_i, i = \overline{1,n}$ are random measurement discrepancy (errors).

To create mathematical models using experimental data for example for monitoring and diagnostic tasks, one has to deal with stochastic heterogeneity. We will point out such features as: incomplete correspondence of some parts of the observations to the model; possible presence of outliers in samplings not necessarily due to measurement errors; often non-experimental, heterogeneous nature of data; use of different groupings and rounding; possible dependence of the observation results [2].

In this case, the use of classical procedures, based on fulfillment of the basic prerequisites of mathematical statistics, can lead to gross estimation errors. In this situation, we use stable (robust and nonparametric) estimation methods based on the least absolute deviation method (LAD) [3], which for model (1) minimizes the sum of the modules of the residuals

$$Q(\mathbf{a}) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{m} a_j x_{ij} \right| = \sum_{i=1}^{n} |y_i - \langle \mathbf{x}_i, \mathbf{a} \rangle| \to \min_{\mathbf{a} \epsilon \mathbf{R}^m}, \tag{2}$$

where $\mathbf{a}=(a_1, a_2, \ldots, a_m)$; $\mathbf{x}_i = (1, x_{i2}, \ldots, x_{im}), i = \overline{1,n}$.

If the outliers are symmetrical, the LAD estimations provide acceptable results. However, unsymmetrical outliers can lead to estimation errors while the least absolute deviation method is using. As an alternative, a generalized method of the least absolute deviations (GLAD) is proposed in [2]. The GLAD estimations for the model (1) are found as a solution to the task:

$$W(\mathbf{a}) = \sum_{i=1}^{n} \rho\left(|y_i - \langle \mathbf{x}_\mathrm{i}, \mathbf{a}\rangle|\right) \to \min_{\mathbf{a}\epsilon\mathbf{R}^m}, \tag{3}$$

where $\mathbf{a}=(a_1, a_2, \ldots, a_m)$; $\mathbf{x}_\mathrm{i} = (1, x_{i2}, \ldots, x_{im})$, $i = \overline{1,n}$.

Computational Experiments and Comparative Analysis

To solve the tasks (2) and (3) in [4] the algorithms based on descent through nodal straight lines are proposed. Computational experiments on model data were carried out in the paper in order to study the effectiveness of the proposed algorithms when estimating the coefficients of multiple linear regression models. In addition, a comparative analysis of the algorithms of descent through nodal straight lines for GLAD and LAD with known exact and approximate methods of solving tasks (2) and (3) was carried out.

The results of the comparison of the algorithm of descent through the nodal straight lines to solve the task (2) with known exact algorithms of solving it (a brute-force search algorithm and the solving of the equivalent problem of linear programming), are reflected in Table. 1 and in Fig. 1.

Table 1

The computational complexity of the algorithms to find the exact solution of the task (2)

Algorithm	Computational complexity
Brute-force search algorithm[2]	$M_1 = O\left(C_n^m \cdot (m^3 + m \cdot n)\right)$
Simplex-method (worst case)[5]	$M_2 = O\left(n^4\right)$
Simplex-method (best case)[5]	$M_3 = O(n^3 \cdot \ln n)$
Descent algorithm through the nodal straight lines	$M_4 = O(m^2 \cdot n^2 + m^4 \cdot n \cdot \ln n + m^2 \cdot n \cdot \ln^2 n)$

In addition, by a statistical testing method, a series of control experiments were conducted to compare the algorithm of descent through the nodal straight lines with an algorithm based on the solving of an equivalent linear programming problem [6]. The results of solving the task (2) where the number of tests $N = 1000$; the number of parameters of the model $m = 5$; the random errors $\boldsymbol{\varepsilon}$ have a distribution in types

$$F(x) = (1-\gamma)\, N(0, \sigma^2) + \gamma \left(\tfrac{1}{\pi} \cdot \mathrm{arctg}\left(\tfrac{x-a_H}{\gamma_H}\right) + \tfrac{1}{2}\right), \tag{4}$$

$\sigma = 1$; $a_H = 0$; $\gamma_H = 1$; $\gamma = 0.1$, are shown in Tables 2 and 3.

In Tables 2 and 3, M is the number of computational operations, $\tilde{M}$ is the average number of computational operations. All confidence intervals for the number of computational operations of the algorithm of descent through the nodal straight lines lie lefter than the intervals for the simplex algorithm. In addition, the simplex algorithm does not always find the exact solution of task (2).

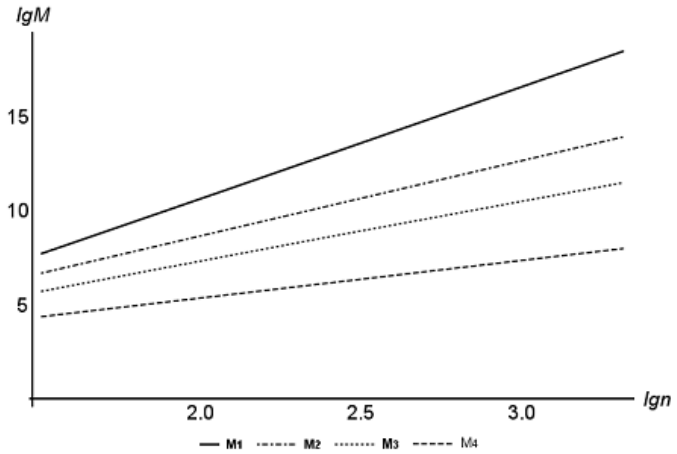

Fig. 1. Computational complexity of the algorithms to find the exact solution of problem (2), $m = 5$

Table 2

The results of solutions of the task (2) with the help of the simplex algorithm

n	Average iteration number	99% confidence interval for the number of iterations		$\lg\tilde{M}$	99%-confidence interval for $\lg M$		The number of discrepancies from the exact solution %
		Left border	Right border		Left border	Right border	
32	20.72	20.25	21.18	4.390	4.380	4.399	5.5
64	40.26	39.45	41.05	5.250	5.241	5.258	7.9
128	75.67	74.18	77.15	6.110	6.101	6.118	8.4
256	143.34	140.65	146.02	6.981	6.973	6.989	9.0
512	343.14	326.40	359.00	7.958	7.937	7.978	10.1

A comparative analysis of the accuracy and performance of the algorithm of descent through nodal straight lines was carried out with approximate algorithms based on the method of iteratively reweighted least squares (Weiszfeld's algorithm) [7] and zero-order iterative optimization methods [8]. The results of statistical tests for model (1) are given in Table. 4, where the random errors have the distribution (4); $N = 1000$ is the number of tests; $m = 5$; $\Delta_1 = \Delta_2 = \cdots = \Delta_m = 1$ is the initial values of steps by coordinate directions; $\mu = 10^{-6}$ is the number to stop the algorithm; $\lambda = 1.5$ is the accelerating factor; $\alpha = 2$ is the step reduction coefficient; $\tau = 1.618$ is the expansion coefficient; $\beta = 0.618$ is the compression ratio; $M_{iter} = 4m$ is the maximum number of failed tests at the current iteration, $t_0 = 1$ is the initial step size, $T = 10^{-6}$ is the minimum step size.

Table 4 reflects $s_v(n, m)$, $s_p(n, m)$ and $s_r(n, m)$ are the standard quadratic deviation of the vector $\widehat{\mathbf{a}}_v$, $\widehat{\mathbf{a}}_p$ and $\widehat{\mathbf{a}}_r$ respectively, which are sample estimates of the coefficients of the multiple linear regression model relative to the vector $\mathbf{a}^*$ of the exact solution of the problem (2), for the Weiszfeld's algorithm; pattern search and adaptive random search methods. $t_v(n, m)$ is the average computation time for the Weiszfeld's algorithm, $t_u(n, m)$ is the average computation time for the algorithm of descent through nodal straight lines.

Table 3

The results of the solutions of problem (2) with the help of the algorithm of descent through the nodal straight lines

n	Average number of node locus considered	99%-confidence interval for node locus considered		$\lg\tilde{M}$	99%-confidence interval for $\lg M$		The number of discrepancies from the exact solution %
		Left border	Right border		Left border	Right border	
32	134.74	131.70	137.78	4.334	4.324	4.343	0
64	241.00	235.70	246.26	4.887	4.877	4.897	0
128	431.86	422.00	441.73	5.442	5.431	5.451	0
256	784.07	765.48	802.65	6.002	5.991	6.012	0
512	1858.45	1766.14	1951.92	6.677	6.655	6.699	0

Table 4

The standard quadratic deviation of the sample estimates of the parameters of model (1), found by the approximate method, with respect to the exact solution of problem (2), $m = 5$

Variationally-weighted quadratic approximations' algorithms for $n = 64$			Methods of searching for an unconditional zero-order extremum		
			n	Configuration Method	Adaptive random search method
$\delta-$computational accuracy	$s_v(n,m)$	$t\,(n,m) = \frac{t_v(n,m)}{t_u(n,m)}$		$s_p(n,m)$	$s_r(n,m)$
10^{-1}	0.632	0.30	32	41.32	38.27
10^{-2}	0.352	0.72	64	44.00	41.21
10^{-3}	0.192	1.56	128	46.91	42.89
10^{-4}	0.127	3.0	256	47.96	44.57
10^{-5}	0.064	4.4	512	51.38	48.36
10^{-6}	0.041	5.6	1024	53.11	50.22

The computational efficiencies of the proposed algorithm of descent through nodal straight lines and the known exact method (a brute-force search algorithm) of solving the task (3) were compared. The results of the comparison are shown in Table 5 and Fig. 2.

Table 5

Computational complexity of algorithms to find the exact solution of the task (3)

Brute-force search algorithm	Descent algorithm through nodal lines for GLAD
$M_1 = O\,(C_n^m \cdot (m^3 + m \cdot n))$	$M_2 = O\,(n^2 \cdot m^3 \cdot lnn + C_{\alpha \cdot n}^m \cdot (m^3 + m \cdot n))$

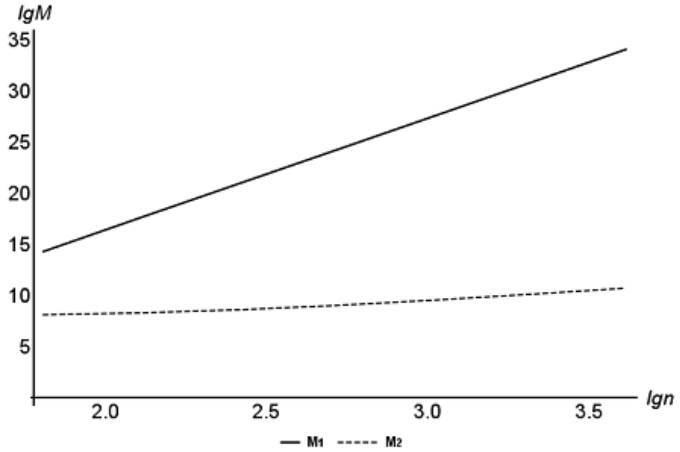

Fig. 2. Graphs of the function of computational complexity of algorithms to find the exact solution of the problem (3)

The results of statistical tests showed that with the increased amount of sampling, the value of the optimal α decreases inversely proportional. And in the case where the errors have a distribution of the form (4) $\alpha = 45/\ n$.

A comparison was also conducted between the algorithm of descent through nodal lines for GLAD and the algorithm of modified GLAD (the algorithm of finding the approximate solution of problem (3)) [9].

The results of statistical tests, where the errors have a distribution of the form (4); $q = 2$ is the number of subsamples; $m = 4$ is the dimension of model; n= 100 is the amount of sampling; N= 1000 is the number of tests, are given in Table 6.

Table 6

The results of comparison of the algorithms of modified GLAD and the algorithm of descent through the nodal straight lines for GLAD

γ	$p^{'}(n,m)$	$t(n,m)/t^{'}(n,m)$	$p^{''}(n,m)$	$t(n,m)/t^{''}(n,m)$
0.05	100	49.3	22.3	20.6
0.1	100	49.3	22	20.6
0.2	99.9	49.3	21.5	20.6

Where $p^{'}(n,m)$ is the percentage of matching of the vector of sample estimates of the coefficients of the multiple linear regression model with the exact solution vector of the problem (3) for the algorithm of descent through the nodal straight lines and $p^{''}(n,m)$ is the same for the modified GLAD, $t^{'}(n,m)$ and $t^{''}(n,m)$ are the average computation time for the algorithm of descent through the nodal straight lines and the modified GLAD algorithm respectively, $t(n,m)$ is the computation time for the brute-force search algorithm.

Conclusion

Based on the results of the analysis it was concluded that the developed new computational algorithms based on descent through the nodal straight lines for realization of the least absolute deviations (LAD) method and the generalized method of the least absolute deviations (GLAD) when estimating the parameters of multiple linear regression models, in terms of computational complexity and accuracy, greatly benefit from known exact and approximate methods and can be effectively used in practice.

References

1. Demidenko E. Z. *Linear and Nonlinear Regression*. Moscow, Finance and Statistics, 1981. (in Russian).
2. Tyrsin A. N. [*Robastic Parametric Identification of Diagnostic Models Based on the Generalized Method of Least Absolute Deviations*]. The Dissertation of the DSc (Techn). Chelyabinsk, South Ural State University, 2007. (in Russian).
3. Mudrov V. I., Kushko V. L. [*Methods of Processing Measurements. Quasi-like Estimations*]. Moscow, Radio and Communications, 1983. (in Russian).
4. Tyrsin A. N., Azaryan A. A. Resistant Linear Model Fitting Methods Based on the Descent Through the Nodal Straight Lines. *Models, Systems, Networks in Economics, Engineering, Nature and Society*, 2018, no. 1, pp. 188–202. (in Russian).
5. Pan V. On the Complexity of a Pivot Step of the Revised Simplex Algorithm. *Computers & Mathematics with Applications*, 1985, vol. 11, no. 11, pp. 1127–1140. doi: 10.1016/0898-1221(85)90190-7.
6. Zukhovitski S. I., Avdeeva L. I. [*Linear and Convex Programming*]. Moscow, Fizmat. Publ., 1967. (in Russian).
7. Akimov P. A., Matasov A. I. Levels of Nonoptimality of the Weiszfeld Algorithm in the Least-Modules Method. *Automation and Remote Control*, 2010, vol. 71, no. 2, pp. 172–184.
8. Panteleev A. V., Letova T. A. [*Optimization Methods in Examples and Tasks*]. Moscow, Vysshaya Shkola Publ., 2002. (in Russian).
9. Tyrsin A. N., Maksimov K. E. [Effective Computational Algorithms for Constructing Regression Models Based on the Generalized Method of Least Absolute Deviations]. *Mathematical Modelling and Boundary Problems*. Proceedings of the Sixth All-Russian Scientific Conference with international participation (1–4 June 2009). Part 4. Information Technology in Mathematical Modelling. Samara, Samara State Technical Univ., 2009, 137–139. (in Russian).

Aleksan A. Azaryan, Faculty of Informatics and Applied Mathematics, Yerevan State University (Yerevan, Republic of Armenia), a.a.azaryan@gmail.com.

Received August 20, 2018.

УДК 519.237.5:519.24 **DOI: 10.14529/jcem180302**

ИССЛЕДОВАНИЕ АЛГОРИТМОВ УСТОЙЧИВОГО ОЦЕНИВАНИЯ КОЭФФИЦИЕНТОВ ЛИНЕЙНЫХ МНОГОМЕРНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

А. А. Азарян

Проведены вычислительные эксперименты на модельных данных с целью исследования эффективности алгоритмов спуска по узловым прямым для оценивания коэффициентов многомерных линейных регрессионных моделей на основе метода наименьших модулей (МНМ) и обобщенного метода наименьших модулей (ОМНМ). Проведен сравнительный анализ данных алгоритмов с известными точными и приближенными методами реализации МНМ и ОМНМ.

Ключевые слова: линейная регрессионная модель; метод наименьших модулей; обобщенный; вычислительная сложность; сравнительный анализ.

Литература

1. Демиденко, Е. З. Линейная и нелинейная регрессия / Е. З. Демиденко. – М.: Финансы и статистика, 1981.
2. Тырсин, А. Н. Робастная параметрическая идентификация моделей диагностики на основе обобщенного метода наименьших модулей: дис. ... д-ра техн. наук / А. Н. Тырсин. – Челябинск, 2007.
3. Мудров, В. И. Методы обработки измерений. Квазиправдоподобные оценки / В. И. Мудров, В. Л. Кушко. – М.: Радио и связь, 1983.
4. Тырсин, А. Н. Методы устойчивого построения линейных моделей на основе спуска по узловым прямым / А. Н. Тырсин, А. А. Азарян // Модели, системы, сети в экономике, технике, природе и обществе. – 2018. – № 1. – С. 188–202.
5. Pan, V. On the Complexity of a Pivot Step of the Revised Simplex Algorithm / V. Pan // Computers & Mathematics with Applications. – 1985. – Vol. 11, № 11. – P. 1127–1140.
6. Зуховицкий, С. И. Линейное и выпуклое программирование / С. И. Зуховицкий, Л. И. Авдеева. – М.: ФИЗМАТЛИТ, 1967.
7. Акимов, П. А. Уровни неоптимальности алгоритма Вейсфельда в методе наименьших модулей / П. А. Акимов, А. И. Матасов // Автоматика и телемеханика. – 2010. – № 2. – С. 4–16.
8. Пантелеев, А. В. Методы оптимизации в примерах и задачах / А. В. Пантелеев, Т. А. Летова. – М.: Высшая школа, 2002.
9. Тырсин, А. Н. Эффективные вычислительные алгоритмы построения регрессионных моделей на основе обобщенного метода наименьших модулей / А. Н. Тырсин, К. Е. Максимов // Математическое моделирование и краевые задачи: труды шестой Всероссийской научной конференции с международным участием (1–4 июня 2009 г.). Часть 4. Информационные технологии в математическом моделировании. – Самара: СамГТУ, 2009. – С. 137–139.

Азарян Алексан Артурович, факультет информатики и прикладной математики, Ереванский государственный университет (г. Ереван, Республика Армения), a.a.azaryan@gmail.com.

Поступила в редакцию 20 августа 2018 г.