Записки научных
семинаров ПОМИ
Том 328, 2005 г.

А. Н. Тырсин

РОБАСТНОЕ ПОСТРОЕНИЕ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ НА ОСНОВЕ ОБОБЩЕННОГО МЕТОДА НАИМЕНЬШИХ МОДУЛЕЙ

1. Введение

При статистической обработке результатов экспериментальных исследований часто решается задача построения линейной регрессионной модели

$$y_i = a_0 + a_1 x_{i1} + \ldots + a_m x_{im} + \epsilon_i, \quad (i = 1, \ldots, n). \tag{1}$$

где $\mathbf{y}^T = (y_1, \ldots, y_n)$ – вектор значений зависимой переменной; $\mathbf{X} = \{x_{ij}\}$, $i = 1, 2, \ldots, n$, $j = 0, 1, \ldots, m$ – матрица значений объясняющих переменных; $\epsilon^T = (\epsilon_1, \ldots, \epsilon_n)$ – вектор случайных ошибок; $\mathbf{a}^T = (a_0, a_1, \ldots, a_m)$ – вектор коэффициентов регрессии.

Среди статистических методов построения регрессионных зависимостей (1) наиболее распространен метод наименьших квадратов (МНК). Он позволяет получить глубокие статистические результаты в предположении, что случайные ошибки ϵ распределены по нормальному закону. Реальные наблюдения во многих задачах в целом соответствуют данной предпосылке. Однако в ряде случаев возникают ситуации, когда предположение о гауссовой модели случайных ошибок не выполняется. При этом использование МНК может привести к значительным ошибкам [1].

С целью обеспечения устойчивости оценок относительно отклонений случайных ошибок от гауссовой модели разработан ряд робастных статистических методов: метод наименьших модулей (МНМ) [2, 3], метод минимакса (M-оценивание) [1, 4], ранговые [5] и знаковый [6] методы. Термин "робастность" означает устойчивость или слабочувствительность статистической процедуры к незначительным отклонениям данных от начальных предположений [1].

Данные методы основаны на более общих предположениях относительно случайных ошибок:
– они независимы и имеют симметричный закон распределения (методы минимакса);

– случайные ошибки независимы и распределены непрерывно и одинаково (ранговые методы);
– независимые случайные ошибки с равными вероятностями принимают положительные и отрицательные значения (МНМ, знаковый метод).

2. Постановка задачи

Наиболее распространенной моделью исследования устойчивости является модель засорения Тьюки–Хьюбера

$$F(x) = (1-\gamma)F_\epsilon(x) + \gamma H(x),$$

где $F_\epsilon(x)$ – функция распределения случайных ошибок ϵ_i, обладающая "хорошими" свойствами (как правило, нормальностью); $H(x)$ – функция распределения засорений, имеющих вид выбросов, как по уровню, так и по дисперсии; γ – вероятность появления выброса.

В теоретических исследованиях обычно в качестве засорения используют случайные величины, у которых математическое ожидание стремится к бесконечности или не существует (распределение Коши). Кроме того, обычно предполагается известной степень засорения γ, что на практике, как правило, невозможно.

Представляется актуальным разработка робастных процедур, удовлетворяющих следующим требованиям [4, 7]:
– они должны быть ориентированы на ограниченное засорение, находящееся "на грани", между "интуитивно возможным" и "интуитивно невозможным";
– степень засорения должна находиться в достаточно широких пределах;
– скорость сходимости выборочных оценок к асимптотическим результатам должна быть достаточно высокой, чтобы обеспечить их практическое применение.

3. Методика

Рассмотрим следующую задачу построения линейной регрессионной модели [8]:

$$(a_0^*, a_1^*, \dots, a_m^*) = \\ = \arg\min_{a_0,\dots,a_m} \sum_{i=1}^{n} \left| y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij} \right|^\lambda, \; 0 < \lambda < 1, \quad (2)$$

где $a_0^*, a_1^*, \ldots, a_m^*$ – искомые параметры.

Данное оценивание имеет преимущество перед МНМ ($\lambda = 1$) и знаковым методом в случае, когда выбросы в ошибках разного знака имеют различные вероятности.

Как известно, целевая функция задачи (2) имеет множество локальных минимумов. Предложенный в [2] метод вариационно-взвешенных квадратических приближений позволяет найти с заданной погрешностью один из локальных минимумов, ближайший к начальному приближению $\mathbf{a}^{(0)}$ вектора $\mathbf{a}$. Глобальный минимум не находится. Более того, идея данного подхода заключается в итерационном приближении решения к некоторому единственному экстремуму, который по умолчанию является глобальным. Этим объясняется, что он нашел применение лишь для реализации метода наименьших модулей.

Теорема 1. *Пусть имеется выборка наблюдений* $(x_1, \ldots, x_n)$ *и задана функция*

$$Q(a) = \sum_{i=1}^{n} |x_i - a|^{\lambda}, \quad 0 < \lambda < 1.$$

Тогда локальными минимумами функции $Q(a)$ *являются все* x_i, *и между любыми двумя соседними точками* x_i *и* x_{i+1} *имеется локальный максимум.*

Доказательство. Определим две первые производные функции $Q(a)$ для всех неособых точек:

$$Q'(a) = \lambda \sum_{i=1}^{n} |x_i - a|^{\lambda - 1} \operatorname{sign}(x_i - a);$$

$$Q''(a) = \lambda(\lambda - 1) \sum_{i=1}^{n} |x_i - a|^{\lambda - 2}.$$

Рассмотрим левый и правый пределы производной $Q'(a)$. Для любого $1 \le i \le n$

$$\lim_{a \to x_i - 0} Q'(a) = -\infty; \qquad \lim_{a \to x_i + 0} Q'(a) = +\infty.$$

Функция $Q(a)$ непрерывна, поэтому все точки x_i разрыва ее первой производной являются локальными минимумами.

Поскольку $0 < \lambda < 1$, то $\forall a \ne x_i : Q''(a) < 0$. Следовательно, для всех значений x_i производная $Q'(a)$ монотонно убывает слева и монотонно возрастает справа от x_i. Поэтому все $(n-1)$ нулей производной $Q'(a)$ соответствуют локальным максимумам, находящимся между точками x_i и x_{i+1}, что и требовалось доказать.

Обобщим теорему 1 на многомерный случай.

Теорема 2. *Пусть имеется выборка наблюдений* $(\mathbf{x}_i, y_i) = (1, x_{i1}, \ldots, x_{im}, y_i)$, $(i = 1, 2, \ldots, n)$, *и задана функция*

$$Q(a_0, \ldots, a_m) = \sum_{i=1}^{n} \left| y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij} \right|^{\lambda}, \quad 0 < \lambda < 1. \tag{3}$$

Введем гиперплоскости $\Omega_i = \Omega(\mathbf{a}, \mathbf{x}_i, y_i)$ *в виде уравнений*

$$y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij} = 0, \quad (i = 1, \ldots, n). \tag{4}$$

Введем также узловые точки пересечения гиперплоскостей (4)

$$\mathbf{u} = \cap_{s \in M} \Omega_s, \quad M = \{k_1, \ldots, k_{m+1}\},$$

$$\forall i \neq j \quad k_i \neq k_j, \quad k = 1, \ldots, n. \tag{5}$$

Обозначим U *– множество всех узловых точек* (5). *Тогда справедливо следующее утверждение.*

Для того чтобы точка $\mathbf{u}^{\circ} = (a_0^{\circ}, \ldots, a_m^{\circ})$ *являлась точкой локального минимума функции* (3) *необходимо и достаточно, чтобы* $\mathbf{u}^{\circ} \in U$.

Доказательство. Очевидно, что все гиперплоскости (4) являются особыми точками функции (3), поскольку в них она не дифференцируема.

Рассмотрим произвольную стационарную точку $\mathbf{a}^{\circ} = (a_0^{\circ}, \ldots, a_m^{\circ})$, для которой $\nabla Q(\mathbf{a}^{\circ}) = 0$. Предположим, что $\mathbf{a}^{\circ}$ является точкой локального минимума. Поскольку эта точка не является особой, то функция (3) в ней дважды дифференцируема.

Тогда матрица Гессе $H(\mathbf{a}^{\circ})$ должна быть неотрицательно определенной, т.е. все ее угловые миноры должны быть неотрицательными [9]:

$$\Delta_1 \geq 0, \ \Delta_2 \geq 0, \ \ldots, \ \Delta_{m+1} \geq 0.$$

Рассмотрим первый угловой минор Δ_1. Поскольку $0 < \lambda < 1$, то для всех неособых точек

$$\Delta_1 = \frac{\partial^2 Q(\mathbf{a}^\circ)}{\partial(\mathbf{a}^\circ)^2} = \lambda(\lambda - 1)\sum_{i=1}^{n}\left|y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij}\right|^{\lambda-2} < 0.$$

Получено противоречие, следовательно, все стационарные точки функции $Q(\mathbf{a})$ не являются локальными минимумами. Таким образом, локальные минимумы, если они существуют, могут находиться только в особых точках.

Рассмотрим q произвольных различных гиперплоскостей Ω_j, $j \in \{j_1, \ldots, j_q\}$, образующих некоторый многогранник W. Очевидно, что $q \geq m + 2$. Покажем, что:

1. Во внутренней части $\operatorname{int}\{W\}$ многогранника W ровно один локальный максимум.

2. Функция Q монотонно убывает по всем направлениям до граничных точек, лежащих на гиперплоскостях Ω_j, $j \in \{j_1, \ldots, j_q\}$.

Во всех внутренних точках $\operatorname{int}\{W\}$ функция (3) непрерывно дифференцируема. Поэтому с учетом доказанного выше в $\operatorname{int}\{W\}$ функция Q не имеет локальных минимумов. Следовательно, у функции (3) не может быть больше одного локального максимума. Предположим теперь, что локальных максимумов у функции Q в $\operatorname{int}\{W\}$ нет. Это возможно только в случае ее линейности. Поскольку функция Q не линейна, то она имеет не менее одного локального максимума.

Следовательно, функция (3) в $\operatorname{int}\{W\}$ имеет ровно один локальный максимум $\mathbf{u}$. В силу непрерывной дифференцируемости функции она будет монотонно убывать при изменении аргумента относительно $\mathbf{u}$ по любым направлениям вплоть до граничных точек многогранника W.

Если рассматривать неограниченный многогранник, то в этом случае локального максимума не будет, т.к. функция (3) будет монотонно возрастать в сторону неограниченной части многогранника.

Рассмотрим некоторую особую точку $\mathbf{a}^k = (a_0^k, \ldots, a_m^k) \in R^{m+1}$, которая удовлетворяет условию $\mathbf{u}^k \in \Omega_k$, k – произвольное наблюдение, $1 \leq k \leq n$. Это равносильно тому, что

$$y_k - a_0^k - \sum_{j=1}^{m} a_j^k x_{kj} = 0.$$

Точка $\mathbf{u}^k$ лежит на границе, как минимум, двух многогранников W_k' и W_k''.

Зададим произвольное направление $\mathbf{z} = (z_0, \ldots, z_m)$, не параллельное гиперплоскости Ω_k, причем $0 < \|\mathbf{z}\| < \infty$.

С учетом доказанного выше, точка $\mathbf{u}^k$ будет локальным минимумом по направлению $\mathbf{z}$ к точке $\mathbf{u}^k$ ввиду монотонного убывания функции (3) по данному направлению для многогранника W_k'. С другой стороны, точка $\mathbf{u}^k$ будет локальным минимумом по направлению $\mathbf{z}$ от точки $\mathbf{u}^k$ для многогранника W_k''.

Таким образом, установили, что все особые точки $\mathbf{u}^k$ являются локальными минимумами по направлениям, не параллельным гиперплоскости Ω_k.

Докажем теперь, что все локальные минимумы функции (3) находятся в узловых точках (5) пересечения гиперплоскостей (4).

Достаточность. Рассмотрим узловую точку $\mathbf{u}^\circ = (a_0^\circ, \ldots, a_m^\circ)$, являющуюся пересечением некоторых $(m+1)$ произвольно взятых гиперплоскостей $\Omega^* = \{\Omega_1, \ldots, \Omega_{m+1}\}$.

Предположим противное, что данная узловая точка $\mathbf{u}^\circ$ не является локальным минимумом. Следовательно, хотя бы на одной из гиперплоскостей $\Omega_k \in \Omega^*$ существует ближайший к точке $\mathbf{u}^\circ$ локальный минимум $\mathbf{u}^k = (a_0^k, \ldots, a_m^k)$, причем $\mathbf{u}^k \notin U$.

Поскольку точка $\mathbf{u}^k$ является локальным минимумом, то она одновременно есть локальный минимум по направлению $\mathbf{z}^k$, параллельному гиперплоскости Ω_k.

Как было показано выше, функция (3) в точке $\mathbf{u}^k$ является непрерывно дифференцируемой по направлению $\mathbf{z}^k$, параллельному гиперплоскости Ω_k. Поэтому точка $\mathbf{u}^k$ является стационарной, и не может быть локальным минимумом, т.к. не является таковым по направлению $\mathbf{z}^k$, параллельному гиперплоскости Ω_k. Следовательно $Q(\mathbf{u}^k) > Q(\mathbf{u}^\circ)$. Получили противоречие.

Поэтому любая узловая точка $\mathbf{u}^\circ$ является локальным минимумом функции (3).

Необходимость. Пусть некоторая точка $\mathbf{u}^k = (a_0^k, \ldots, a_m^k)$ гиперплоскости Ω_k является локальным минимумом функции (3). Предположим, что эта точка не является узловой.

Согласно доказанному выше, функция $Q(\mathbf{u})$ в точке $\mathbf{u}^k$ не может быть локальным минимумом по направлению $\mathbf{z}^k$, параллельному гиперплоскости Ω_k, т.е. получили противоречие. Следовательно, локальные минимумы функции (3) лежат в узловых точках.

Таким образом, доказали, что $\mathbf{u}^k = (a_0^k, \ldots, a_m^k)$ – локальный минимум функции (3) тогда и только тогда, когда $\mathbf{u}^k$ является узловой точкой.

Замечание 1. Теорема 2 позволяет достаточно просто найти глобальный минимум задачи (2). Поиск сводится к перебору точек пересечения гиперплоскостей (5).

Замечание 2. МНМ-оценки называют также оценками нормы L_1 или L_1-оценками [3]. По аналогии ОМНМ-оценки (2) назовем L_p-оценками, $0 < p < 1$.

Замечание 3. Очевидно, что ОМНМ может применяться и для вычисления коэффициентов авторегрессионных моделей $y_k = a_1 y_{k-1} + \ldots + a_p y_{k-p} + \epsilon_k$.

При этом решается задача

$$(a_1^*, \ldots, a_p^*) = \arg \min_{a_1, \ldots, a_p} \sum_{k=p+1}^{n} \left| y_k - \sum_{j=1}^{p} a_j y_{k-j} \right|^{\lambda}.$$

Решение задачи (2) обладает рядом полезных свойств. Во-первых, в отличие от МНМ, большие ошибки здесь уменьшаются. Во-вторых, L_p-оценки менее чувствительны к несимметричным выбросам по сравнению с L_1-оценками.

Задачу (2) можно обобщить на следующий случай. Очевидно, что функция потерь

$$f(x) = |x|^{\lambda} \tag{6}$$

при $0 < \lambda < 1$ является вогнутой на положительной полуоси.

Оказывается, что использование вместо (6) любой монотонно возрастающей, дважды непрерывно-дифференцируемой на положительной полуоси и равной нулю при $x = 0$ вогнутой функции сохраняет верными теоремы 1, 2.

Теорема 3. *Пусть имеется выборка наблюдений* $(\mathbf{x}_i, y_i) = (1, x_{i1}, \ldots, x_{im}, y_i)$, $(i = 1, 2, \ldots, n)$, *и задана функция*

$$Q(a_0, \ldots, a_m) = \sum_{i=1}^{n} g\left(\left| y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij} \right|\right), \tag{7}$$

где $g(x)$ – некоторая монотонно возрастающая, всюду дважды непрерывно-дифференцируемая на положительной полуоси функция, причем $g(0) = 0$, $g''(x) < 0$.

Для того чтобы точка $\mathbf{u}^\circ = (a_0^\circ, \dots, a_m^\circ)$ являлась точкой локального минимума функции (3) необходимо и достаточно, чтобы она принадлежала множеству узловых точек пересечения гиперплоскостей (5), т.е. $\mathbf{u}^\circ \in U$.

Доказательство. Определим знак первого углового минора Δ_1. В силу вогнутости функции g для всех неособых точек

$$\Delta_1 = \frac{\partial^2 Q(\mathbf{a}^\circ)}{\partial(\mathbf{a}^\circ)^2} = \sum_{i=1}^{n} g'' \left(\left| y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij} \right| \right) < 0.$$

Все остальное доказательство полностью совпадает с доказательством теоремы 2.

В классе функций, удовлетворяющих условиям теоремы 3, представляет интерес функция потерь

$$g(x) = \arctan |x|. \tag{8}$$

При этом решается задача

$$(a_1^*, \dots, a_m^*) = \arg \min_{a_0, \dots, a_m} \sum_{i=1}^{n} \arctan \left(\left| y_i - a_0 - \sum_{j=1}^{m} a_j x_{ij} \right| \right). \tag{9}$$

где $a_0^*, a_1^*, \dots, a_m^*$ – искомые параметры.

Функция (8) обладает следующими особенностями.

Во-первых, она асимптотически ограничена. Иными словами, оценки (9) более устойчивы к наличию в измеренных данных сверхбольших выбросов по сравнению с оценками (2).

Во-вторых, при малых погрешностях измерений (если $\sigma_\epsilon^2 \to 0$) дисперсия оценок (2) будет значительно превосходить дисперсию оценок (9). Действительно, при $0 < \lambda < 1$, $\lim_{x\to+0}(|x|^\lambda)' = +\infty$, в то время как $\lim_{x\to+0}(\arctan |x|)' = 1$.

В-третьих, на основе функции (9) можно построить квазиоптимальные оценки

$$g(x) = \arctan\left(|x|^\tau\right), \quad \tau > 1, \quad \tau \neq 2l, \quad l \in N, \tag{10}$$

сочетающие устойчивость по отношению к выбросам с низкой дисперсией при малых погрешностях. Это можно объяснить тем, что функция (10) для τ вблизи 2,43 при $0 \leq x \leq 1$ с точностью до постоянного множителя близка к параболе, соответствующей МНК.

Можно отметить следующие преимущества оценок на основе функции потерь (10) по сравнению с M-оценками [1, 4]:
– у функции (10) отсутствует параметр сдвига, который нужно задавать априорно;
– решение носит прямой характер, а не итерационный.

Квазиоптимальность оценок вызвана тем, что при $\tau > 1$ функция (10) является вогнутой при $x > w$ и выпуклой, если $0 < x < w$, $[(\tau-1)/(\tau+1)]^{1/2\tau}$. Поэтому в общем случае функция (10) может иметь локальные минимумы в стационарных точках. Однако область выпуклости ограничена сверху ($w < 1$), причем $w \to 1$ при $\tau \to \infty$.

Замечание 4. Оценки с функцией потерь (10) для удобства назовем $L_{g(\tau)}$-оценками.

Приведенные выше результаты по робастному оцениванию можно обобщить на случай нелинейных регрессионных моделей вида

$$y_i = a_0 + f(\mathbf{x}_i; a_1, \ldots, a_m) + \epsilon_i,$$

где $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$; f – некоторая всюду непрерывно дифференцируемая в области определения функция m переменных $a_1, \ldots, a_m$.

Теорема 4. *Пусть имеем задачу*

$$(a_1^*, \ldots, a_m^*) = \arg \min_{a_0, \ldots, a_m} \sum_{i=1}^{n} g(|y_i - a_0 - f(\mathbf{x}_i; a_1, \ldots, a_m)|). \quad (11)$$

Все локальные минимумы задачи (11) *являются решениями систем уравнений*

$$\begin{cases} a_0 + f(\mathbf{x}_{i1}; a_1, \ldots, a_m) = y_{i1}, \\ \ldots \\ a_0 + f(\mathbf{x}_{i,m+1}; a_1, \ldots, a_m) = y_{i,m+1}, \end{cases} \quad (12)$$

где $i_1, \ldots, i_{m+1} \in \{1, \ldots, n\}$, $\forall k \neq l \ i_k \neq i_l$.

Доказательство теоремы 4 аналогично доказательству теоремы 2.

Замечание 5. Глобальный минимум задачи (11) находится путем перебора всех решений системы (12).

4. Экспериментальная часть

Проведем сравнительный анализ оценок методом Монте-Карло. В данной статье ограничимся случаем конечных выборок приемлемого объема с точки зрения практики.

Случай 1. Исследование устойчивости оценок относительно выбросов.

Рассмотрим модель

$$y_i = \alpha + \epsilon_i,\ (i = 1, \dots, n), \tag{13}$$

где случайные ошибки ϵ_i имеют распределение

$$F(x) = (1-\gamma)N(0,\sigma^2) + \gamma N(\mu,\sigma_1^2), \tag{14}$$

представляющее собой смесь двух нормальных распределений.

Очевидно, распределение (14) является симметричным при $\mu = 0$. При $\sigma_1 > \sigma$ имеем случай, соответствующий удлинению хвостов распределений.

Для определенности зададим: $\alpha = 0$; $\sigma = 1$; $\mu = 2$; $\sigma_1 = 3$; $n = 200$ – объем выборки; $M = 20000$ – число испытаний. Случайные ошибки с распределением (14) моделируем методом, описанным в [10]. Результаты расчетов представлены в таблицах 1 и 2. Здесь приведены не сами оценки, а их абсолютные погрешности Δa и дисперсии s_a^2, оценки a параметра модели (13).

Таблица 1

Абсолютные погрешности Δa оценок параметра сдвига α

γ	L_2	L_1	$L_{0,75}$	$L_{0,5}$	$L_{0,25}$	$L_{g(1)}$	$L_{g(1,5)}$	$L_{g(2,5)}$
0	0,0085	0,0121	0,0100	0,0105	0,0092	0,0089	0,0134	0,0222
0,025	0,0614	0,0238	0,0171	0,0121	0,0062	0,0027	0,0123	0,0208
0,05	0,1106	0,0377	0,0274	0,0187	0,0129	0,0084	0,0165	0,0247
0,1	0,2033	0,0738	0,0537	0,0400	0,0309	0,0220	0,0285	0,0355
0,15	0,3119	0,1072	0,0746	0,0550	0,0421	0,0238	0,0292	0,0339
0,2	0,4097	0,1450	0,1046	0,0813	0,0604	0,0420	0,0472	0,0483
0,25	0,4975	0,2045	0,1493	0,1157	0,1008	0,0725	0,0717	0,0712
0,3	0,5959	0,2552	0,1854	0,1389	0,1127	0,0769	0,0775	0,0742

Таблица 2

Выборочные дисперсии s_a^2 оценок параметра сдвига α

γ	L_2	L_1	$L_{0,75}$	$L_{0,5}$	$L_{0,25}$	$L_{g(1)}$	$L_{g(1,5)}$	$L_{g(2,5)}$
0	0,0083	0,0147	0,0218	0,0311	0,0419	0,0314	0,0238	0,0200
0,025	0,0088	0,0153	0,0226	0,0332	0,0466	0,0344	0,0259	0,0216
0,05	0,0108	0,0155	0,0222	0,0325	0,0462	0,0328	0,0255	0,0222
0,1	0,0123	0,0152	0,0229	0,0339	0,0545	0,0347	0,0267	0,0233
0,15	0,0186	0,0171	0,0242	0,0364	0,0569	0,0373	0,0275	0,0232
0,2	0,0277	0,0182	0,0243	0,0391	0,0594	0,0367	0,0243	0,0208
0,25	0,0351	0,0248	0,0311	0,0446	0,0695	0,0422	0,0270	0,0217
0,3	0,0384	0,0298	0,0376	0,0524	0,0762	0,0472	0,0293	0,0224

Для наглядности на рис. 1 и 2 приведем соответствующие графики результатов оценивания.

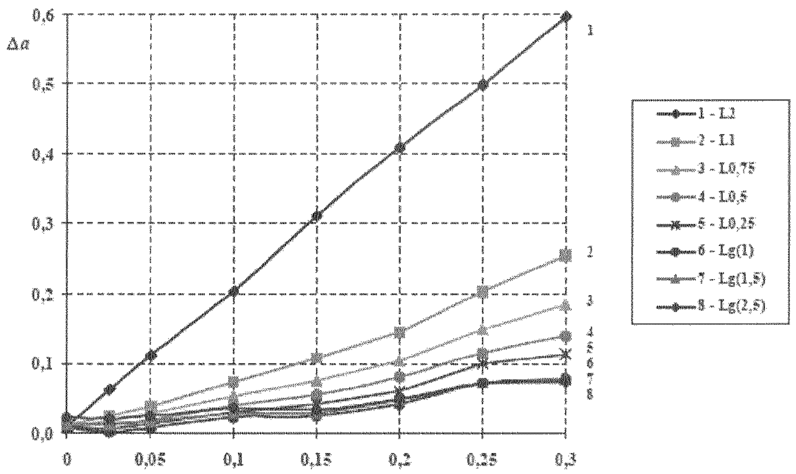

Рис. 1. Абсолютные погрешности Δa оценок параметра сдвига α.

Оценим погрешность выборочных результатов оценивания. Поскольку имеем собственно-случайную повторную выборку, то предельная ошибка оценки a вследствие ограниченности выборки равна $\Delta a = ts_a/\sqrt{M}$, где t – табличное значение коэффициента доверия, соответствующее заданной доверительной вероятности.

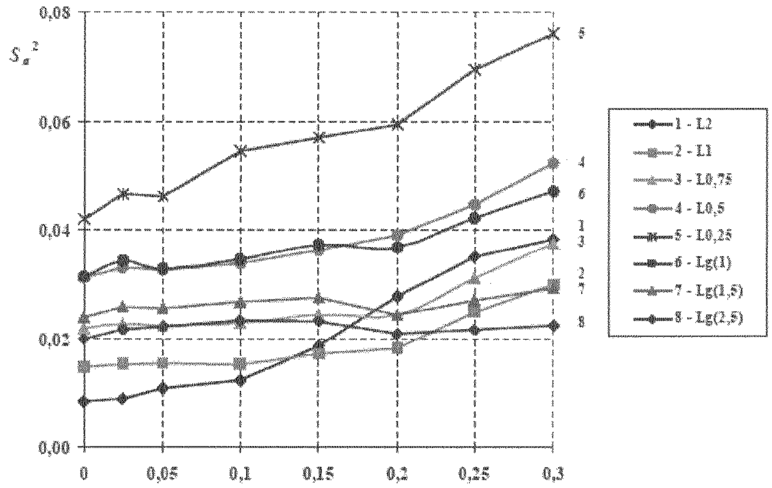

Рис. 2. Выборочные дисперсии s_a^2 оценок параметра сдвига α.

Для определенности примем доверительную вероятность равной 0,954. Предельные ошибки оценок вследствие ограниченности выборки ведут себя достаточно стабильно и составляют в среднем 0,0024. Значения предельных ошибок оценок приведены в таблице 3.

Таблица 3

Предельные ошибки оценок a

γ	L_2	L_1	$L_{0,75}$	$L_{0,5}$	$L_{0,25}$	$L_{g(1)}$	$L_{g(1,5)}$	$L_{g(2,5)}$
0	0,0013	0,0017	0,0021	0,0025	0,0029	0,0025	0,0022	0,0020
0,025	0,0013	0,0017	0,0021	0,0026	0,0031	0,0026	0,0023	0,0021
0,05	0,0015	0,0018	0,0021	0,0026	0,0030	0,0026	0,0023	0,0021
0,1	0,0016	0,0017	0,0021	0,0026	0,0033	0,0026	0,0023	0,0022
0,15	0,0019	0,0018	0,0022	0,0027	0,0034	0,0027	0,0023	0,0022
0,2	0,0024	0,0019	0,0022	0,0028	0,0034	0,0027	0,0022	0,0020
0,25	0,0027	0,0022	0,0025	0,0030	0,0037	0,0029	0,0023	0,0021
0,3	0,0028	0,0024	0,0027	0,0032	0,0039	0,0031	0,0024	0,0021

Случай 2. Исследование устойчивости оценок относительно гетероскедастичности ошибок.

Рассмотрим модель (13), в которой случайные ошибки ϵ_i имеют распределение $F_\epsilon(x) = N(0, \sigma^2)$, дисперсия которого зависит от некоторого фактора пропорциональности t, т.е. $\sigma^2 = \sigma^2(t)$. В качестве фактора пропорциональности t используем номер анализируемого значения в выборке, $t = i$.

Рассмотрим следующие виды гетероскедастичности: линейная модель $\sigma_1(t) = c_1 t\sigma$; параболическая модель: $\sigma_2(t) = c_2(t-\tau_1)^2\sigma$; экспоненциальная модель: $\sigma_3(t) = c_3 \exp(t/\tau_2)\sigma$.

Для определенности зададим: $\alpha = 0$; $\sigma = 1$; $n = 50$ – объем выборки; $M = 50000$ – число испытаний; $c_1 = 0,2$; $c_2 = 0,025$; $c_3 = 0,2$; $\tau_1 = 25$; $\tau_2 = 12$. В качестве засоряющей модели используем модель (14), в которой $\gamma = 0,1$.

Результаты оценивания различными методами представлены в таблицах 4 и 5.

Таблица 4

Абсолютные погрешности Δa оценок параметра сдвига α при гетероскедастичности

Вид гетероскедастичности		L_p-оценки				$L_{g(\tau)}$-оценки	
		p=2	p=1	p=0,75	p=0,5	τ=1	τ=3
Линейная	без засорения	0,1007	0,0793	0,0735	0,0681	0,0443	0,0328
	с 10%-м засореннем	1,0281	0,1452	0,0729	0,0362	0,0127	0,0144
Параболическая	без засорения	0,0954	0,0276	0,0130	0,0056	0,0072	0,0180
	с 10%-м засорением	1,1030	0,0511	0,0142	0,0039	0,0006	0,0015
Экспоненциальная	без засорения	0,0648	0,0224	0,0205	0,0193	0,0165	0,0149
	с 10%-м засорением	0,6097	0,0329	0,0085	0,0021	0,0091	0,0038

Таблица 5

Выборочные дмсперсии s_a^2 оценок параметра сдвига α при гетероскедастичности

Вид гетероскедастичности		L_p-оценки				$L_{g(\tau)}$-оценки	
		p=2	p=1	p=0,75	p=0,5	τ=1	τ=3
Линейная	без засорения	0,7922	0,3015	0,2975	0,3321	0,4469	0,4866
	с 10%-м засорением	1,7130	0,3760	0,3552	0,3955	0,5333	0,5921
Параболическая	без засорения	1,1106	0,0404	0,0127	0,0052	0,0060	0,0527
	с 10%-м засорением	2,2317	0,0516	0,0155	0,0058	0,0063	0,0525
Экспоненциальная	без засорения	0,5159	0,0338	0,0302	0,0326	0,0268	0,0322
	с 10%-м засорением	1,0677	0,0407	0,0338	0,0364	0,0301	0,0352

Значения предельных ошибок оценок вследствие ограниченности выборки приведены в таблице 6.

Таблица 6

Предельные ошибки оценок a при гетероскедастичности

Вид гетероскедастичности		L_p-оценки				$L_{g(\tau)}$-оценки	
		p=2	p=1	p=0,75	p=0,5	τ=1	τ=3
Линейная	без засорения	0,0080	0,0049	0,0049	0,0052	0,0060	0,0062
	с 10%-м засорением	0,0117	0,0055	0,0053	0,0056	0,0065	0,0069
Параболическая	без засорения	0,0094	0,0018	0,0010	0,0006	0,0007	0,0021
	с 10%-м засорением	0,0134	0,0020	0,0011	0,0007	0,0007	0,0020
Экспоненциальная	без засорения	0,0064	0,0016	0,0016	0,0016	0,0015	0,0016
	с 10%-м засорением	0,0092	0,0018	0,0016	0,0017	0,0016	0,0017

Полученные оценки предельных ошибок оценок вследствие ограниченности выборки в обоих случаях оказались незначительны. Следовательно, результаты расчетов в целом характеризуют свойства рассмотренных оценок.

5. Выводы, направления дальнейших исследований

1. Предложен новый метод устойчивого оценивания линейных моделей – обобщенный метод наименьших модулей. Данный метод обобщает МНМ со случая L_1 на L_p, $0 < p < 1$, и далее на более общие функции потерь. Доказаны необходимые и достаточные условия для нахождения точного решения.

2. L_p-оценки имеют более высокую устойчивость при одностороннем засорении и в случае гетероскедастичности ошибок по сравнению с L_1-оценками. Одновременно они сохраняют устойчивость и при симметричном засорении.

3. Получены квазиоптимальные оценки, сочетающие устойчивость к выбросам с достаточно низкой дисперсией.

4. Установлено, что при некоторых не столь существенных ограничениях, данный подход можно распространить и на случай нелинейных регрессионных моделей.

5. Представляется целесообразным дальнейшие исследования ОМНК проводить в следующих направлениях [11]:

– необходимо исследовать устойчивость оценок на основе функций влияния [12];

– исследовать асимптотические свойства предложенного метода;

– разработать точные (не асимптотические) статистические тесты (построение статистических тестов и доверительных множеств).

Литература

1. П. Хьюбер, *Робастность в статистике.* Мир, М. (1984).
2. В. И. Мудров, В. Л. Кушко, *Методы обработки измерений. Квазиправдоподобные оценки.* — Радио и связь, М. (1983).
3. Y. Dodge (ed.), *Statistical data analysis based on the L_1-norm and related methods.* — in Papers of the 4th International Conference on Statistical Analysis on the L_1-Norm and Related Methods, Basel, Birkhauser (2002), 454 p.
4. M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, *Theory and Applications of Recent Robust Methods, Statistics for Industry and Technology.* Басел, Биркхаусер (2004).
5. Т. Хеттманспергер, *Статистические выводы, основанные на рангах.* — Финансы и статистика, М. (1987).
6. М. В. Болдин, Г. И. Симонова, Ю. Н. Тюрин, *Знаковый статистический анализ линейных моделей.* — Наука М. (1997).
7. А. И. Орлов, *Эконометрика.* — Экзамен, М. (2002).
8. A. M. Tyrsin, *Robust Identification of Linear Models on Experimental Data.* — in The 4th Moscow International Conference On Operations Research: Proceedings, Moscow (2004), pp. 221–223.
9. Э. М. Галеев, *Оптимизация: теория, примеры, задачи.* — Эдиториал УРСС, М. (2002).
10. Дж. Форсайт, М. Малькольм, К. Моулер, *Машинные методы математических вычислений.* — Мир, М. (1980).
11. R. Dutter, P. Filzmoser, U. Gather, and P. J. Rousseeuw, Developments in Robust Statistics. — Heidelberg-New York, Physica Verlag (200).
12. Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль, *Робастность в статистике. Подход на основе функций влияния.* — Мир, М. (1989).

Tyrsin A. N. Robust regression models definition on base of generalised method of least modules.

In the paper, a new method of steady construction of linear regression models, namely, the generalized method of the least modules is offered. The method is theoretically justifed. Experimental approbation on the test data is considered.

Челябинский государственный университет

Поступило 7 октября 2005 г.