Вісник Харківського національного університету
Серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»
№ 847, 2009, с.276-291

УДК 519.233.5

Робастное оценивание параметров полиномиальной регрессии второго порядка

А. В. Омельченко
Харьковский национальный университет радиоэлектроники, Украина

It was described a new approach to robust estimates of polynomial regression parameters construction. This technique is based on forming a set of elemental estimates we obtain by the least squares method with a following finding of a sample median of the given set. Robust estimates of polynomial regression of the second degree parameters have been constructed with a sequential analysis of them. The major attention is paid to analysis of the robust estimates of the leading coefficient of a polynomial regression of the second degree. It is shown that we can gain the robustness of estimation in comparison with the known methods.

Введение

Многие задачи анализа данных и обработки сигналов решаются с использованием полиномиальных моделей. Для оценивания параметров этих моделей традиционно используется метод наименьших квадратов (МНК). Однако оценки МНК не являются устойчивыми к нарушениям модельных предположений и наличию у распределений, описывающих ошибки наблюдений, более тяжелых, по сравнению с нормальным распределением, «хвостов».

В 60-х годах прошлого столетия в работах Хьюбера была заложена теория робастного оценивания, в частности робастного регрессионного анализа. Согласно классификации Хьюбера [1] существуют следующие робастные оценки: M-оценки, L-оценки и R-оценки. M-оценки являются оценками максимального правдоподобия, L-оценки строятся на основе линейных комбинаций порядковых статистик, а R-оценки – на основе ранговых статистик.

Наиболее глубокие результаты в теории робастного оценивания были получены для одномерного случая [1-3]. Для робастного регрессионного анализа предложен ряд алгоритмов оценивания, общим недостатком которых является их итерационный характер, затрудняющий анализ свойств оценок [3-6].

В настоящее время в связи с прогрессом вычислительной техники появилась возможность реализации новых робастных алгоритмов оценивания, обладающих высокой вычислительной сложностью, но простой структурой, позволяющей проводить анализ свойств таких оценок.

Целью настоящей работы является разработка нового подхода к построению робастных оценок параметров полиномиальной регрессии и исследованию их свойств. Этот подход основывается на формировании множества оценок, получаемых МНК, с последующим нахождением выборочной медианы для данного множества. Основное внимание уделено оцениванию коэффициента при старшей степени полиномиальной регрессии второго порядка.

Для достижения сформулированной цели в работе решаются следующие основные задачи.

1. Разрабатываются робастные оценки параметров полиномиальной регрессии второго порядка.
2. Исследуются свойства разработанных оценок.
3. Выполняется имитационное моделирование, подтверждающее эффективность разработанных оценок.

1. Постановка задачи

Будем полагать, что наблюдается последовательность случайных величин $\{(\tau_k, z_k):\ k=\overline{0,n-1}\}$ таких, что

$$z_k = \sum_{p=0}^{m} a_p \cdot \tau_k^p + \xi_k, \quad k=\overline{0,n-1}, \tag{1}$$

где a_p, $p=\overline{0,m}$ – постоянные величины; ξ_k, $k=\overline{0,n-1}$ – последовательность независимых случайных величин; $m=2$.

Следуя [3], будем считать, что векторы $(\tau_k,\xi_k), k=\overline{0,n-1}$ статистически независимы и имеют функцию распределения

$$Q(t,x)=(1-\varepsilon)\cdot F(t,x)+\varepsilon\cdot H(t,x),\ \ k=\overline{0,n-1}, \tag{2}$$

где ε – вероятность засорения; $F(t,x)=K(t)\cdot\Phi(x)$ – двумерная функция распределения пары случайных величин (τ_k,ξ_k) при отсутствии засорения; $H(t,x)$ – функция распределения засорения, принадлежащая некоторому классу Ψ. Полагается, что при отсутствии засорения случайные величины ξ_k подчиняются нормальному закону с функцией распределения $\Phi(x)$ и имеют нулевое математическое ожидание и дисперсию σ^2, а случайные величины τ_k, $k=\overline{0,n-1}$ обладают функцией распределения $K(t)$.

Требуется найти устойчивые оценки $\hat{a}_p$ коэффициентов a_p, $p=\overline{0,m}$.

2. Показатели качества робастных оценок

Обозначим символами Q, F, H вероятностные распределения, которые описываются соответствующими функциями распределения $Q(t,x)$, $F(t,x)$, $H(t,x)$.

Пусть в качестве оценок параметров a_p, $p\in\overline{0,m}$ используются статистики $S_n^p[(\tau_0,z_0),...,(\tau_{n-1},z_{n-1})]$, обозначаемые далее как S_n^p, которые при $n\to\infty$ сходятся по вероятности к соответствующим величинам $S^p(Q)$.

На основе результатов работы [3] функцию влияния для оценки S_n^p определим как предел

$$IF_p(t,x) = \lim_{\varepsilon \to 0+} \left\{ \frac{S^p((1-\varepsilon)F + \varepsilon G_{t,x}) - S^p(F)}{\varepsilon} \right\}, \quad (3)$$

где $G_{t,x}$ – двумерное распределение, сосредоточенное в точке (t,x).

Асимптотические дисперсии оценок S_n^p определяются через функцию влияния [3]

$$V_p = \int [IF_p(t,x)]^2 dF(t,x), \ p = \overline{0,m}. \quad (4)$$

Чувствительность оценки S_n^p к большой ошибке определяется как [3]

$$\gamma_p^* = \sup_{t,x} \left| IF_p(t,x) \right|. \quad (5)$$

Если γ_p^* конечна, то соответствующая оценка S_n^p называется B-робастной. Оценки, которые не допускают одновременного уменьшения по дисперсиям V_p и по чувствительности γ_p^* называются оптимальными B-робастными оценками.

Асимптотическую пороговую точку оценки S_n^p определим следующим образом [3]:

$$\varepsilon_p^* = \sup\{\varepsilon : \exists \ r_\varepsilon > 0 \ \text{такое, что при} \ n \to \infty$$

$$\text{из} \ Q \in \{(1-\varepsilon)F + \varepsilon H, H \in \Psi\} \ \text{следует} \ P\{|S_n^p| \le r_\varepsilon\} \to 1 \ \}. \quad (6)$$

Пороговая точка указывает наибольшую долю резко выделяющихся наблюдений, при которой оценка еще не становится бесполезной.

Кроме асимптотической пороговой точки существует определение пороговой точки при конечной выборке [3]. Пороговой точкой оценки S_n^p при конечной выборке $(z_0,...,z_{n-1})$ называется величина, определяемая из условия

$$\varepsilon_p^*(n) = \frac{1}{n} \max\{q : \max_{i_1,...,i_q} \sup_{y_1,...,y_q} \left| S_n^p(u_0,...,u_{n-1}) \right| < \infty\}, \quad (7)$$

где выборка $(u_0,...,u_{n-1})$ получена заменой q точек данных $z_{i_1},...,z_{i_q}$ произвольными значениями $y_1,...,y_q$.

3. M-оценки параметров регрессии

M-оценки параметров регрессии минимизируют выражение

$$\sum_{k=0}^{n-1} \rho(z_k - \sum_{p=0}^{m} a_p \cdot t_k^p), \quad (8)$$

где $\rho(\cdot)$ – некоторая функция, которая обычно является выпуклой. Взяв производную выражения (8) по параметрам a_p, $p = \overline{0,m}$, придем к системе уравнений

$$\sum_{k=0}^{n-1} \psi(z_k - \sum_{j=0}^{m} a_j \cdot t_k^j) \cdot t_k^p = 0, \ p = \overline{0,m}, \quad (9)$$

где $\psi(z) = \rho'(z)$. Функция $\psi(\cdot)$ обычно нелинейная, поэтому М-оценки, как правило, не сохраняют свойства инвариантности масштаба, как это имеет место в методе наименьших квадратов. Для того, чтобы выполнялось свойство инвариантности оценок параметров регрессии к масштабу минимизируют выражение

$$\sum_{k=0}^{n-1} \rho\{(z_k - \sum_{p=0}^{m} a_p \cdot t_k^p)/s\},$$

где s – помехоустойчивая оценка параметра масштаба, зависящая от остаточных разностей

$$e_k = z_k - \sum_{p=0}^{m} a_p \cdot t_k^p, \qquad k = \overline{0, n-1}.$$

В качестве такой оценки часто используется медиана абсолютных отклонений (MAD) [2, 3]

$$s = 1{,}483 \cdot \mathrm{med}\left|e_k - \mathrm{med}\ e_k\right|.$$

В литературе предложено несколько видов функции ψ: функция Хьюбера; функция усеченного среднего, функция Хампеля, функция Андрюса, биквадратная функция Тьюки [1-3].

Наиболее изученной и широко используемой является функция Хьюбера, имеющая следующий вид:

$$\psi(x) = \begin{cases} x, & \text{если } |x| < r; \\ r \cdot \mathrm{sign}(x), & \text{если } |x| \ge r. \end{cases} \tag{10}$$

Для одномерного случая ($p = 0$) доказано, что функция (10) является оптимальной в смысле минимаксного критерия [1].

Функция усеченного среднего описывается выражением

$$\psi(x) = \begin{cases} x, & \text{если } |x| < r; \\ 0, & \text{если } |x| \ge r, \end{cases} \tag{11}$$

а функция Хампеля

$$\psi(x) = \begin{cases} x, & \text{если } |x| < b; \\ b \cdot \mathrm{sign}(x), & \text{если } b \le |x| < c; \\ b \cdot \dfrac{r - |x|}{r - c} \mathrm{sign}(x), & \text{если } c \le |x| < r; \\ 0, & \text{если } |x| \ge r, \end{cases} \tag{12}$$

где $0 < b \le c < r < \infty$.

М-оценки, соответствующие функциям (11) и (12), относятся к классу сниженных оценок, для которого $\psi(x) = 0$ при $|x| \ge r$. Сниженные М-оценки полезны, когда априори известно, что засоряющее распределение H в основном сосредоточено вне интервала $(-r, r)$, т. е по отношению к резко выделяющимся наблюдениям. Недостатком таких оценок является их чувствительность к неверной оценке масштаба и вычислительные проблемы, связанные с не

единственностью решения системы уравнений (9) [3]. Кроме того, немонотонный характер функции $\psi(x)$ не должен быть слишком резким, иначе возможны проблемы со сходимостью алгоритмов оценивания [2].

В [3] показано, что вид функции влияния М-оценок $IF(t,x)$ при любом t будет определяться видом функций $\psi(x)$.

4. Начальные оценки

Для построения М-оценок параметров регрессии (а также R-оценок) требуется начальная оценка. Начальная оценка имеет исключительное значение, так как она определяет свойства М-оценок [2]. Особенно чувствительны к ее выбору оценки, полученные на основе немонотонных функций ψ. При неудачном начальном приближении итерационный процесс может сходиться к локальному минимуму. Кроме того, даже в случае сходимости неудачный выбор начального приближения может потребовать слишком большого числа итерационных циклов. Обычно в качестве начального приближения используется оценка метода наименьших квадратов. Однако такая оценка сильно зависит от исходных данных. Хорошая начальная оценка коэффициентов регрессии сама должна быть робастной. Поэтому предложен ряд методов для вычисления устойчивой начальной оценки [2]. Одним из основных требований к таким начальным оценкам должно быть требование высокой пороговой точки. Величина дисперсии для начальной оценки имеет меньшее значение, поскольку эта оценка подлежит дальнейшему уточнению.

В качестве начальной оценки могут использоваться [2]: метод Тейла, метод, использующий коэффициенты Спирмэна, и ортогональный метод Брауна-Муда. Все указанные методы являются итерационными. Наиболее простым и эффективным из описанных методов является метод Тейла.

Модифицированный метод Тейла описан в [2]. В нем для получения ортогональных переменных применяется процедура ортогонализации Грама-Шмидта, после чего линейная модель (1) принимает вид

$$z_k = \sum_{p=0}^{m} \alpha_p \cdot x_{k,p} + \xi_k, \qquad k = \overline{0, n-1}. \tag{13}$$

Далее находятся коэффициенты регрессии α_p, $p = \overline{1,m}$ с использованием итерационной процедуры следующего вида:

$$\begin{aligned}
&1)\ d_p(k,s) = (z_s - z_k)/(x_{s,p} - x_{k,p}),\ s > k,\ k = \overline{0, n-2};\\
&2)\ \delta\alpha_p = \operatorname*{med}_{k,s}\ d_p(k,s);\\
&3)\ \alpha_p \leftarrow \alpha_p + \delta\alpha_p;\\
&4)\ z_k \leftarrow z_k - \delta\alpha_p x_{k,p},\ k = \overline{0, n-1}.
\end{aligned} \tag{14}$$

В каждом цикле процедуры шаги 1-4 повторяются последовательно для переменных α_p, $p = \overline{1,m}$. Процедура начинается со значений $\alpha_p = 0$, $p = \overline{1,m}$ и продолжается до тех пор, пока не будет достигнут нужный уровень сходимости. После этого определяется оценка коэффициента

$$\hat{\alpha}_0 = \text{med } z_k . \tag{15}$$

Для возврата к оценкам начальных коэффициентов регрессии, применяется преобразование, обратное к использованному преобразованию Грама-Шмидта.

К недостаткам алгоритма Тейла относится его реализационная сложность, связанная с большим объемом вычислений, а также трудность анализа свойств получаемых оценок, обусловленная итерационным характером процедуры вычисления коэффициентов и нелинейностью операций в каждой из таких итераций.

5. Робастные оценки параметров полиномиальной регрессии второго порядка, использующие медианную обработку элементарных оценок МНК

Построим алгоритмы оценивания коэффициента при старшей степени a_2 для модели (1) на основе совокупности статистик (элементарных оценок)

$$\theta(\tau_i, \tau_j, \tau_k) = \frac{1}{\tau_k - \tau_i} \cdot \left[\frac{z_k - z_j}{\tau_k - \tau_j} - \frac{z_j - z_i}{\tau_j - \tau_i} \right], \tag{16}$$

каждая из которых является оценкой МНК, найденной по трем точкам: (τ_i, z_i), (τ_j, z_j), (τ_k, z_k). Для обеспечения устойчивости результирующих оценок используем два подхода к их формированию: в первом из них оценку параметра a_2 определим как выборочную медиану во всем множестве статистик (16), а во втором используем метод повторяющихся медиан.

В первом подходе оценку коэффициента a_2 зададим как выборочную медиану статистик (16)

$$\hat{a}_2 = \underset{\tau_i < \tau_j < \tau_k}{\text{med}} \ \theta(\tau_i, \tau_j, \tau_k) . \tag{17}$$

Оценки коэффициентов a_1 и a_0 определим путем последовательного понижения порядка модели:

$$\hat{a}_1 = \underset{\tau_i < \tau_j}{\text{med}} \theta(\tau_i, \tau_j) ,$$

$$\hat{a}_0 = \underset{i}{\text{med}} \ \breve{\breve{z}}_i ,$$

где

$$\theta(\tau_i, \tau_j) = \frac{\breve{z}_j - \breve{z}_i}{\tau_j - \tau_i} , \tag{18}$$

$$\breve{z}_i = z_i - \hat{a}_2 \cdot \tau_i^2 , \quad \breve{\breve{z}}_i = \breve{z}_i - \hat{a}_1 \cdot \tau_i , \quad i = \overline{0, n-1} . \tag{19}$$

Во втором подходе для формирования результирующей оценки по совокупности статистик (16) используем метод повторяющихся медиан. Оценку коэффициента a_2 зададим следующим образом

$$\hat{a}_2 = \underset{\tau_i}{\text{med}} \{ \underset{\tau_j \neq \tau_i}{\text{med}} \{ \underset{\tau_k \neq \tau_i, \tau_j}{\text{med}} \ \theta(\tau_i, \tau_j, \tau_k) \} \} , \tag{20}$$

где статистики $\theta(\tau_i, \tau_j, \tau_k)$ определяются согласно (16).

Оценки коэффициентов a_1 и a_0 также найдем в результате понижения порядка модели:

$$\hat{a}_1 = \underset{\tau_i}{\text{med}}\{\underset{\tau_j \neq \tau_i}{\text{med}}\ \theta(\tau_i, \tau_j)\}, \qquad \hat{a}_0 = \underset{i}{\text{med}}\ \breve{\breve{z}}_i,$$

где статистики $\theta(\tau_i, \tau_j)$, $\breve{z}_i$, $\breve{\breve{z}}_i$ определяются согласно выражениям (18, 19).

6. Анализ свойств оценок

Оценки (17) и (20) обладают рядом полезных свойств.

Утверждение 1. Если для наблюдаемых данных справедлива модель (1), то показатели точности (смещение и дисперсия) оценок (17) и (20) не зависят от истинных значений коэффициентов a_0, a_1, a_2.

Доказательство этого утверждения следует из инвариантности статистики $\theta(t_i, t_j, t_k) - a_2$, определяемой согласно (16) для любых попарно различных τ_i, τ_j, τ_k, к значениям коэффициентов a_0, a_1, a_2 в модели (1).

Данное свойство позволяет изучать свойства оценок параметров модели (1), произвольным образом задав значения коэффициентов a_0, a_1, a_2.

Утверждение 2. Если для наблюдаемых данных справедлива модель (1) и шум наблюдения ξ_k, $k = \overline{0, n-1}$ представляет собой последовательность независимых случайных величин, имеющих симметричное распределение, то оценки (17) и (20) являются несмещенными.

Доказательство данного утверждения выполним в следующей последовательности. Согласно утв. 1 положим $a_0 = a_1 = a_2 = 0$. Тогда

$$\theta(\tau_i, \tau_j, \tau_k) = \frac{\xi_k}{(\tau_k - \tau_i)(\tau_k - \tau_j)} + \frac{\xi_j}{(\tau_j - \tau_i)(\tau_j - \tau_k)} + \frac{\xi_i}{(\tau_i - \tau_j)(\tau_i - \tau_k)}, \qquad (21)$$

где ξ_i, ξ_j, ξ_k – независимые случайные величины.

Изменим знак перед каждой из величин ξ_k, $k = \overline{0, n-1}$. В результате получим последовательность случайных величин $\xi'_k = -\xi_k$, $k = \overline{0, n-1}$. Тогда согласно (21) придем к совокупности решающих статистик

$$\theta'(\tau_i, \tau_j, \tau_k) = -\theta(\tau_i, \tau_j, \tau_k), \qquad \tau_i, \tau_j, \tau_k \in [0, T].$$

Использование этих статистик в алгоритмах (17) или (20) приведет к изменению знака перед оценкой $\hat{a}'_2 = -\hat{a}_2$. Поэтому

$$M[\hat{a}'_2] = -M[\hat{a}_2]. \qquad (22)$$

С другой стороны, в силу статистической независимости и симметрии распределения случайных величин ξ_k, $k = \overline{0, n-1}$ справедливо равенство

$$M[\hat{a}'_2] = M[\hat{a}_2]. \qquad (23)$$

Согласно (22) и (23) получим $M[\hat{a}_2] = 0$, что доказывает несмещенность оценки $\hat{a}_2$, определяемой согласно (17) или (20).

Утверждение 3. Если величины τ_k в модели (1) принимают значения $t_0,...,t_{N-1}$ с одинаковыми вероятностями $p_i = 1/N$, $i = 0,...,N-1$, то функция влияния оценки (17) определяется выражением

$$IF_2(t_i, x) = 3\frac{0{,}5 - H_{i,x}(0)}{f_\theta(0)}, \tag{24}$$

где

$$f_\theta(0) = \frac{1}{C_N^3\sqrt{2\pi}} \sum_{\substack{i,j,k=0 \\ t_i<t_j<t_k}}^{N-1} D^{-\frac{1}{2}}(t_i, t_j, t_k); \tag{25}$$

$$D(t_i, t_j, t_k) = \frac{\sigma^2}{(t_i - t_j)^2(t_i - t_k)^2} + \frac{\sigma^2}{(t_j - t_i)^2(t_j - t_k)^2} + \frac{\sigma^2}{(t_k - t_i)^2(t_k - t_j)^2}; \tag{26}$$

$$H_{i,x}(0) = \frac{1}{C_{N-1}^2} \sum_{\substack{j,k=0 \\ t_j<t_k \\ t_j,t_k\neq t_i}}^{N-1} \Phi(0, \mu_{i,j,k}(x), \sigma_{i,j,k}); \tag{27}$$

$\Phi(x, \mu, \sigma)$ – функция распределения гауссовской величины с математическим ожиданием μ и дисперсией σ^2;

$$\mu_{i,j,k}(x) = \frac{x}{(t_i - t_j)(t_i - t_k)}; \ \sigma_{i,j,k} = \sigma\sqrt{\frac{1}{(t_k - t_j)^2(t_k - t_i)^2} + \frac{1}{(t_j - t_k)^2(t_j - t_i)^2}}. \tag{28}$$

Доказательство. Поскольку статистика $\theta(t_i, t_j, t_k) - a_2$ инвариантна значениям регрессионных коэффициентов, то функция влияния оценки (17) не зависит от истинных значений этих коэффициентов. Поэтому без ограничения общности будем полагать $a_0 = a_1 = a_2 = 0$.

В отсутствие засорения распределение совокупности решающих статистик $\theta(\tau_i, \tau_j, \tau_k)$, $0 \le \tau_i < \tau_j < \tau_k \le T$ представляет собой смесь с плотностью вероятности

$$f_\theta(y) = \frac{1}{C_N^3} \sum_{\substack{i,j,k=0 \\ t_i<t_j<t_k}}^{N-1} \frac{1}{\sqrt{2\pi D(t_i, t_j, t_k)}} \exp\{-\frac{y^2}{2D(t_i, t_j, t_k)}\},$$

где дисперсии статистик $\theta(\tau_i, \tau_j, \tau_k)$ определяются выражением (26).

Предположим, что при выполнении события $\{\tau_k = t_i\}$ в отсчет помехи ξ_k с вероятностью $N \cdot \varepsilon$ вносится засорение со значением x. В этом случае вероятность засорения в точке (t_i, x) равна ε. Это означает, что доля засоренных статистик $\theta(\tau_i, \tau_j, \tau_k)$ составит

$$\rho = \frac{C_{N-1}^2 \varepsilon N}{C_N^3} = 3\varepsilon. \tag{29}$$

Функция распределения засорения для статистики θ имеет вид

$$H_{i,x}(y) = \frac{1}{C^2_{N-1}} \sum_{\substack{j,k=0 \\ t_j<t_k \\ t_j,t_k \neq t_i}}^{N-1} \Phi(y, \mu_{i,j,k}(x), \sigma_{i,j,k}).$$

Будем считать, что при $n \to \infty$ оценка $\hat{a}_2$ сходится по вероятности к величине S. Поскольку согласно (17) оценка $\hat{a}_2$ определяется как выборочная медиана множества статистик θ, то функцию влияния найдем как предел

$$IF_2(t_i, x) = \lim_{\varepsilon \to 0+} \left\{ \frac{S((1-\rho)F_\theta + \rho H_{i,x}) - S(F_\theta)}{\varepsilon} \right\}, \qquad (30)$$

где F_θ – распределение статистики θ при отсутствии засорения, $H_{i,x}$ – распределение статистики θ с учетом засорения.

Известно [1], что для медианы

$$\lim_{\rho \to 0+} \left\{ \frac{S((1-\rho)F_\theta + \rho H_{i,x}) - S(F_\theta)}{\rho} \right\} = \frac{0{,}5 - H_{i,x}(0)}{f_\theta(0)}. \qquad (31)$$

Из (31) с учетом (30) и (29) получим искомое выражение (24) для функции влияния.

Несложно показать, что функции влияния (24) симметричны относительно начала координат, т. е. удовлетворяют условию $IF(t_i, -x) = -IF(t_i, x)$ для всех $x \in R$.

Формулы (24-28) позволяют вычислять значения функций влияния. В качестве иллюстрации на рис. 1 показан вид функции влияния для следующего случая: $N = 20$; $\sigma = 1$; $t_i = i$, $i = \overline{0,19}$. При этом учтена симметричность функции относительно начала координат.

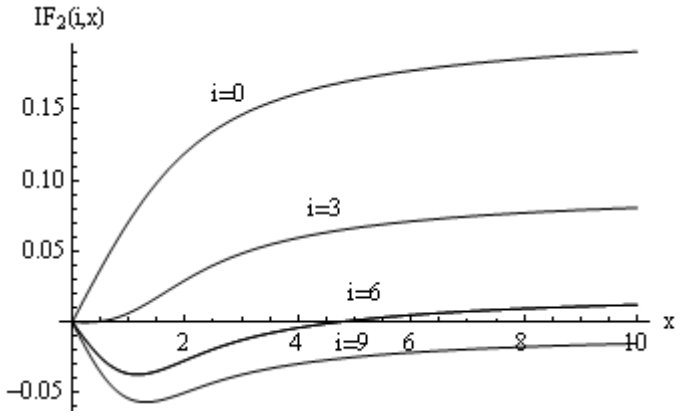

Рис. 1 – Вид функций влияния оценки (17)

Утверждение 4. Если выполняются условия утв. 3, то асимптотическая дисперсия оценки (17) определяется по формуле

$$V_2(N) = \frac{1}{N}\sum_{i=0}^{N-1}\int_{-\infty}^{\infty}[IF_2(t_i, x)]^2 d\Phi(x). \quad (32)$$

Доказательство. Доказательство данного утверждения вытекает из соотношения (4), в котором распределение величин τ_k является сосредоточенным в точках t_i, $i = \overline{0, N-1}$. В результате придем к выражению (32).

Можно показать, что если выполняются условия утв. 3, то асимптотическая дисперсия оценки МНК определяется выражением

$$V_{2\ МНК}(N) = \frac{N\sigma^2}{\sum_{k=0}^{N-1}\varphi^2(t_k)}, \quad (33)$$

где функция

$$\varphi(t) = (t-\bar{t})^2 - \alpha\cdot(t-\bar{t}) - \beta,$$

$$\bar{t} = \frac{1}{N}\sum_{k=0}^{N-1} t_k; \quad \alpha = \frac{\sum_{k=0}^{N-1}(t_k-\bar{t})^3}{\sum_{k=0}^{N-1}(t_k-\bar{t})^2}; \quad \beta = \frac{1}{N}\sum_{k=0}^{n-1}(t_k-\bar{t})^2.$$

Важным показателем эффективности робастных оценок является коэффициент асимптотической эффективности робастной оценки относительно оценки МНК

$$k_a(N) = \frac{V_{2\,МНК}(N)}{V_2(N)}. \quad (34)$$

Рассчитанная согласно формулам (32-34) зависимость коэффициента асимптотической эффективности оценки (17) от числа N представлена на рис. 2.

Утверждение 5. Если выполняются условия утв. 3 и длина интервала наблюдения $T < \infty$, то оценка (17) является B-робастной. При этом чувствительность к большой ошибке

$$\gamma_2^* = \sup_{t,x}\left|IF_2(t,x)\right| \le \frac{3}{2\cdot f_\theta(0)}. \quad (35)$$

Доказательство. Из выражения (24) следует неравенство (35). В то же время, если $T < \infty$, то согласно (25) $f_\theta(0) > 0$. Поэтому $\gamma^* < \infty$ и, следовательно, оценка (17) является B-робастной.

Утверждение 6. Если в модели (1, 2) величины τ_k, $k = \overline{0, n-1}$ имеют непрерывное распределение, а засорение вносится только в величины ξ_k, то асимптотическая пороговая точка оценки (17) удовлетворяет условию $\varepsilon^* = 1 - \sqrt[3]{1/2} \approx 0{,}206$.

Доказательство. Пусть вероятность засорения величин ξ_k, $k = \overline{0, n-1}$ равна ε. Тогда вероятность не засорения статистики $\theta(\tau_i, \tau_j, \tau_k)$, вычисленной согласно (16), равна $(1-\varepsilon)^3$. Если $(1-\varepsilon)^3 > 1/2$, то общая доля засоренных статистик θ будет меньше пороговой точки медианы в (17). Поэтому $\varepsilon < \varepsilon^*$, где величина $\varepsilon^* = 1 - \sqrt[3]{1/2}$ является верхней гранью множества тех величин ε, для которых существует такое конечное число $r_\varepsilon > 0$, что при $n \to \infty$ имеет место сходимость

$$P\{|\hat{a}_2| \le r_\varepsilon\} \to 1. \tag{36}$$

Покажем, что ε^* – точная верхняя грань множества тех ε, для которых имеет место сходимость (36). Для этого рассмотрим некоторое распределение Q_ε векторов (τ_k, ξ_k) в модели (1, 2), удовлетворяющее условию утв. 6. Для этого распределения дисперсию шума ξ_k в отсутствие засорения положим равной нулю. Будем считать, что в ξ_k засорение вносится тогда и только тогда, когда $\tau_k > t_\varepsilon$, где константа t_ε удовлетворяет условию $P(\tau_k > t_\varepsilon) = \varepsilon$. В качестве засорения используем величины $\xi_k = A \cdot (\tau_k - t_\varepsilon)^2$, где $A >> 1$. Без ограничения общности будем полагать, что коэффициенты регрессии $a_0 = a_1 = a_2 = 0$. В этом случае статистика $\theta(\tau_i, \tau_j, \tau_k) = 0$, если $\tau_i, \tau_j, \tau_k < t_\varepsilon$, и $\theta(\tau_i, \tau_j, \tau_k) >> 0$ в противном случае. Очевидно, что если $\varepsilon < \varepsilon^*$, то согласно (17) $\hat{a}_2 = 0$, если же $\varepsilon > \varepsilon^*$, то $\hat{a}_2 > 0$ и при достаточно большом числе A условие (36) не выполняется для любого конечного числа r_ε. Из этого следует, что ε^* является точной верхней гранью множества тех ε, для которого имеет место сходимость (36). Следовательно ε^* – пороговая точка оценки (17).

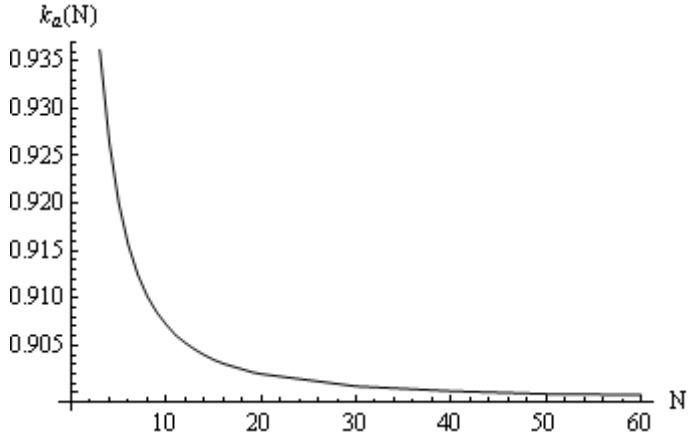

Рис. 2 – Зависимость коэффициента асимптотической эффективности оценки (17) от числа N

Утверждение 7. Если $\tau_k \neq \tau_j$ при $k \neq j$ для всех $k, j = \overline{0, n-1}$, то пороговая точка оценки (17) удовлетворяет условию

$$\varepsilon^*(n) = 1 - \frac{K_n}{n}, \tag{37}$$

где

$$K_n = \min_k \{k : \frac{k(k-1)(k-2)}{n(n-1)(n-2)} > \frac{1}{2}\}. \tag{38}$$

Доказательство. Очевидно, что при общем числе наблюдений n число всех возможных вариантов статистик $\theta(i_1, i_2, i_3)$, $0 \le i_1 < i_2 < i_3 \le n-1$ (17) равно $C_n^3 = n(n-1)(n-2)/6$. Если среди n наблюдений не засоренными оказываются k наблюдений, то общее число всех не засоренных статистик $\theta(i_1, i_2, i_3)$ равно $C_k^3 = k(k-1)(k-2)/6$. Если же при этом отношение $\frac{C_k^3}{C_n^3} = \frac{k(k-1)(k-2)}{n(n-1)(n-2)} > \frac{1}{2}$, то доля засоренных статистик $\theta(i_1, i_2, i_3)$, $0 \le i_1 < i_2 < i_3 \le n-1$ меньше пороговой точки выборочной медианы. Поэтому $k > K_n$, а число засоренных наблюдений $q = n - k$ должно удовлетворять условию: $q < n - K_n$. Аналогично утв. 6 можно показать, что если $q > n - K_n$ то существует такое распределение векторов (τ_k, ξ_k), что условие (7) нарушается. Поэтому пороговая точка оценки (17) определяется равенством (37).

Отметим, что согласно (37) $\lim\limits_{n\to\infty} \varepsilon^*(n) = 1 - \sqrt[3]{1/2} \approx 0{,}206$.

7. Исследование робастных алгоритмов методом имитационного моделирования

Сравнение оценок коэффициентов квадратичной регрессии выполним по коэффициенту эффективности

$$k_e = D_{МНК} / R,$$

где $D_{МНК}$ – дисперсия оценки наименьших квадратов; $R = M[\hat{a}_2 - a_2]^2$ – среднее значение квадрата отклонения робастной оценки $\hat{a}_2$ относительно истинного значения.

В процессе моделирования значение показателя R находилось по 10000 реализаций сигналов. При моделировании принималось $n = N$. Поскольку показатели точности рассматриваемых оценок не зависят от истинных коэффициентов полинома (1), то при моделировании использовались значения $a_0 = a_1 = a_2 = 0$.

На рис. 3 приведены зависимости значений коэффициента эффективности k_e робастных оценок, полученных методом имитационного моделирования, от числа наблюдаемых отсчетов N. Здесь крестиками отображены результаты моделирования для алгоритма (17), квадратиками – для алгоритма (20) и

кружками − для алгоритма Тейла. Пунктирной линией на рис. 3 показано значение коэффициента асимптотической эффективности алгоритма (17), рассчитанное согласно (32-34) при числе отсчетов $N = 50$.

Из результатов моделирования, представленных на рис. 3, следует, что при числе отсчетов $10 \le N \le 30$ разброс оценок приближенно может быть определен по формулам: для алгоритма (17) и алгоритма Тейла $R \approx 1{,}2 \cdot D_{МНК}$, а для алгоритма (20) $R \approx 1{,}5 \cdot D_{МНК}$.

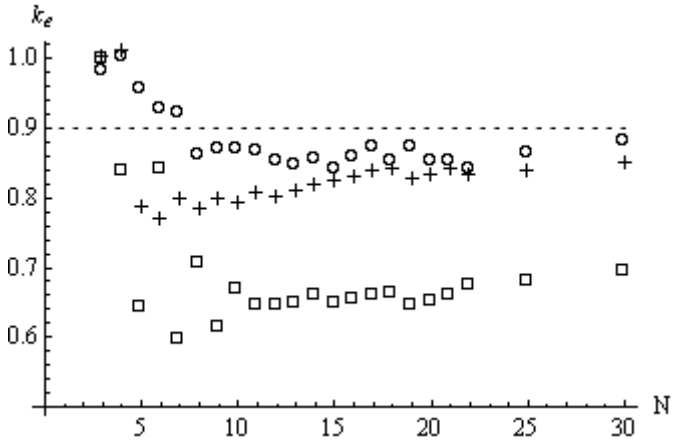

Рис. 3 – Зависимости коэффициента эффективности оценок от числа наблюдаемых отсчетов

Методом моделирования построена зависимость разброса оценок R от числа засоренных отсчетов q. При этом также использовалась модель (1, 2) со следующими параметрами: $n = N = 20$; $t_i = i$, $i = \overline{0,19}$; $a_0 = a_1 = a_2 = 0$; $\sigma = 1$. В качестве засорения использовался шум с нулевым средним и среднеквадратическим отклонением $\sigma = 100$. При этом засорение вносилось в q случайным образом выбранных отсчетов. Показатель R определялся методом имитационного моделирования по 10000 реализаций сигналов.

На рис. 4 – 7 точками отображены результаты моделирования для оценок МНК, а также начальных оценок (17), (20) и оценки Тейла. На этих же рисунках кружочками отображены зависимости разброса R для М-оценок с усеченной функцией ψ вида (11) при соответствующем выборе начальных оценок. Значение порога в (11) $r = 1{,}5$.

Из результатов моделирования следует, что оценка МНК является очень чувствительной к влиянию засорения и при $q \ge 2$ не удается на ее основе получить М-оценку из-за проблем, возникающих с обращением матриц. Начальные оценки (17), (20) и Тейла являются устойчивыми к влиянию засорения, причем наиболее стойкой является оценка (20). Переход от начальных оценок к М-оценкам позволяет повысить устойчивость оценивания коэффициента a_2 полиномиальной регрессии. Причем наиболее стойкими

являются те М-оценки, в которых в качестве начальных используются оценки (20), а наименее − М-оценки, использующие начальные оценки Тейла.

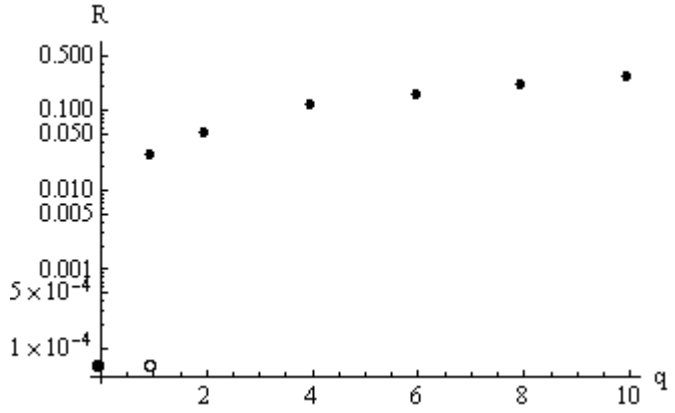

Рис. 4 − Зависимость разброса оценки МНК от числа засоренных отсчетов

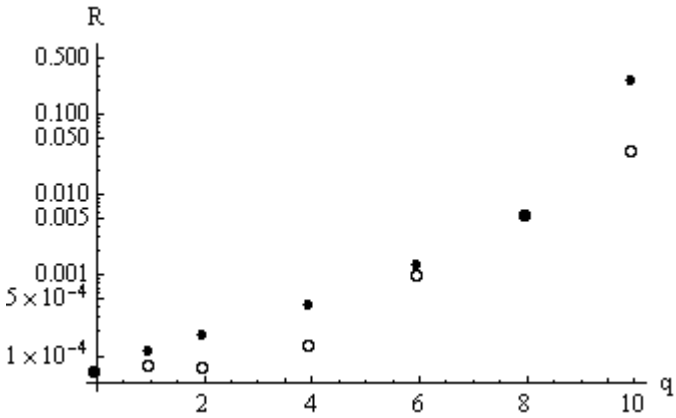

Рис. 5 − Зависимость разброса оценки (17) от числа засоренных отсчетов

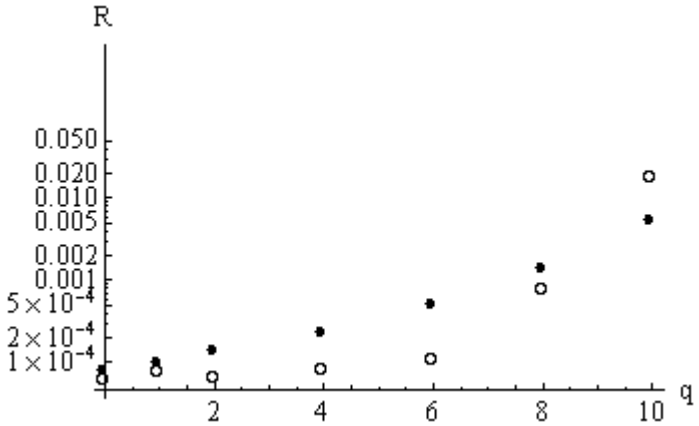

Рис. 6 − Зависимость разброса оценки (20) от числа засоренных отсчетов

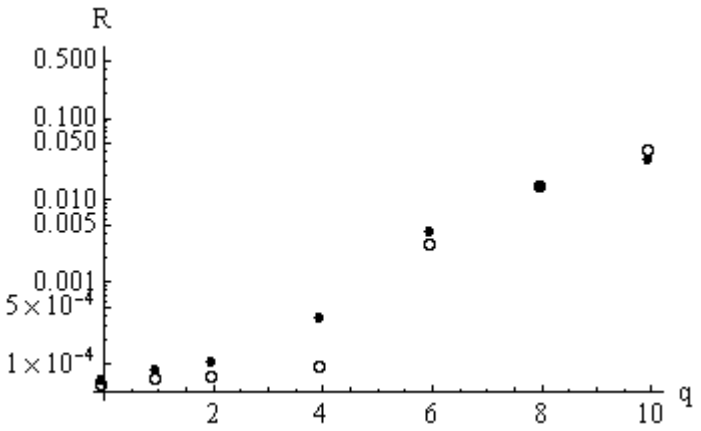

Рис. 7 − Зависимость разброса оценки Тейла от числа засоренных отсчетов

В результате моделирования было также установлено, что применение М-оценок с ψ -функцией Хьюбера (10) существенно ухудшает стойкость оценок к большим уровням загрязнения по сравнению с усеченной ψ -функцией (11).

Выводы

В настоящей работе предложен новый подход к построению робастных оценок параметров полиномиальной регрессии второго порядка, который основывается на формировании множества элементарных оценок, получаемых МНК по трем наблюдениям, с последующим нахождением выборочной медианы для данного множества. Основное внимание уделено оцениванию коэффициента при старшей степени для полиномиальной модели второго порядка.

В рамках предложенного подхода построены два алгоритма оценивания, один из которых основан на поиске выборочной медианы во всем множестве элементарных оценок, а второй использует идею метода повторяющихся медиан.

Выполнен анализ свойств построенных оценок. Установлено, что эти оценки имеют высокий коэффициент асимптотической эффективности относительно оценок МНК, который для оценки (17) достигает значения 0,9. Доказана B-робастность оценки (17) и определена ее асимптотическая пороговая точка $\varepsilon^* = 1 - \sqrt[3]{1/2}$. Показано, что предложенные алгоритмы позволяют повысить устойчивость оценивания по сравнению с методом Тейла.

Методом имитационного моделирования установлено, что применение разработанных оценок в качестве начальных для М-оценок позволяет обеспечить большую устойчивость, чем применение для этой цели оценок Тейла.

Ввиду высокой эффективности оценок (17) при выполнении модельных предположений и их стойкости к засорению выборок они могут быть рекомендованы к использованию в ситуациях, когда объем вычислительных затрат не является критическим.

Дальнейшие исследования планируется посвятить теоретическому исследованию оценок (20), реализующих метод повторяющихся медиан, и решению прикладных задач регрессионного анализа на основе разработанного подхода.

ЛИТЕРАТУРА

1. Хьюбер Дж. П. Робастность в статистике: пер. с англ. − М.: Мир, 1984. – 304 с.
2. Устойчивые статистические методы оценки данных/ Пер. с англ. Ю.И. Малахова под ред. Н.Г. Волкова. − М.: Машиностроение, 1984. – 232 с.
3. Робастность в статистике. Подход на основе функций влияния: Пер. с англ./ Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. − М.: Мир, 1989. – 512 с.
4. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов. − М.: Горячая линия−Телеком, 2007. – 522 с.
5. Вучков Л., Бояджиева Л., Солаков Е. Прикладной линейный регрессионный анализ. − М.: Финансы и статистика, 1987. – 238 с.
6. Olive D.J. Applied Robust Statistics [on-line resource]. – Carbondale, IL USA: Southern Illinois University, Department of Mathematics, 2008. - 588 p. - Access mode: www.math.siu.edu/olive/ol-bookp.htm

Надійшла у першій редакції 29.03.2009, в останній – 08.04.2009.