РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

АВТОМЕТРИЯ

2008, том 44, № 6

АНАЛИЗ И СИНТЕЗ СИГНАЛОВ И ИЗОБРАЖЕНИЙ

УДК 519.233.5

ОЦЕНИВАНИЕ ПАРАМЕТРОВ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ С ИСПОЛЬЗОВАНИЕМ АППРОКСИМАЦИИ ГРАМА – ШАРЛЬЕ *

В. И. Денисов, В. С. Тимофеев

Новосибирский государственный технический университет, г. Новосибирск
E-mail: videnis@nstu.ru

Рассмотрена задача оценивания параметров регрессионных моделей. Предложен новый метод оценивания параметров для распределений случайной ошибки, отличающихся от нормального и представимых в виде ряда Грама – Шарлье типа А. С помощью вычислительных экспериментов проведено исследование разработанного метода в сравнении с традиционным методом наименьших квадратов и знаковым методом.

Введение. Количество информации, привлекаемой для прикладного статистического анализа, оказывает непосредственное влияние на точность результатов (оценок) и качество выводов. Знание первого момента (математического ожидания) позволяет говорить о средних величинах, в том числе и о регрессиях (регрессия определяется как условное математическое ожидание) [1, 2]. Предельные теоремы в этом случае помогают делать выводы о точности получаемых результатов [2]. Второй момент (дисперсия) дает возможность учесть неоднородность условий проведения наблюдений, связанную, например, с влиянием неучтенных в модели факторов либо с использованием других измерителей (приборов), имеющих отличающиеся настройки или обеспечивающих несколько другую точность измерений. Третий и четвертый моменты либо определяемые на их основе коэффициенты асимметрии и эксцесса несут в себе информацию о форме распределения. Известно, что знание коэффициента асимметрии при унимодальности распределения позволяет уточнить неравенство Чебышева [3].

В данной работе сделана попытка оценивания параметров регрессионных зависимостей с использованием коэффициентов асимметрии и эксцесса.

* Работа выполнена при поддержке Министерства образования и науки РФ (проект РНП.2.1.2.43).

Постановка задачи и основные предположения. Рассмотрим регрессионное уравнение вида

$$y = X\theta + \varepsilon, \tag{1}$$

где

$$X = \begin{bmatrix} f_1(x_{11}) & \cdots & f_p(x_{1p}) \\ \vdots & \ddots & \vdots \\ f_1(x_{N1}) & \cdots & f_p(x_{Np}) \end{bmatrix}$$

– матрица значений регрессионных функций, имеющая полный столбцовый ранг, т. е. $rg(X) = p$; $\theta = (\theta_1, ..., \theta_p)^T$ – вектор неизвестных параметров, подлежащих оцениванию; p – количество неизвестных параметров; N – количество проведенных экспериментов; $f(x) = (f_1(x), ..., f_p(x))^T$ – вектор известных действительных функций; x_{ij} – заданные значения входных факторов в N наблюдениях; $y = (y_1, ..., y_N)^T$ – вектор значений отклика; $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)^T$ – вектор ошибок наблюдений.

Будем предполагать, что ошибки ε_i наблюдений являются независимыми одинаково распределенными случайными величинами с плотностью $\varphi(x)$, для которых верно

$$E(\varepsilon_i) = 0, \quad D(\varepsilon_i) = \sigma^2.$$

Также будем предполагать, что существуют третий и четвертый центральные моменты.

Аппроксимация Грама – Шарлье. Следуя [2, 4], отметим, что при условии существования требуемых моментов рассматриваемой случайной величины плотность $\varphi(x)$ может быть представлена в виде ряда по производным некоторой эталонной функции $\varphi_0(x)$. В качестве такой функции обычно выступает функция плотности стандартного нормального распределения

$$\varphi_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Запишем несколько первых производных данной функции:

$$\frac{d}{dx}\varphi_0(x) = -x\frac{1}{\sqrt{2\pi}}e^{-x^2/2} = -x\varphi_0(x), \qquad \frac{d^2}{dx^2}\varphi_0(x) = (x^2 - 1)\varphi_0(x),$$

$$\frac{d^3}{dx^3}\varphi_0(x) = (3x - x^3)\varphi_0(x), \qquad \frac{d^4}{dx^4}\varphi_0(x) = (x^4 - 6x^2 + 3)\varphi_0(x).$$

Очевидно, что результаты дифференцирования представляют собой полиномы от x, умноженные на исходную функцию $\varphi_0(x)$, т. е. так называемые

полиномы Чебышева – Эрмита $H_r(x)$, определяемые в общем случае тождеством [4]

$$\frac{d^r}{dx^r}\varphi_0(x) = (-1)^r H_r(x)\varphi_0(x) \text{ или } H_r(x) = x^r + \sum_{i=1}^{[r/2]} (-1)^i (2i-1)!! C_r^{2i} x^{r-2i},$$

при этом очевидно, что $H_0(x) = 1$. Кроме того, полиномы Чебышева – Эрмита обладают важным условием ортогональности

$$\int_{-\infty}^{\infty} H_k(x) H_l(x) \varphi_0(x) dx = \begin{cases} 0, & k \neq l, \\ k!, & k = l. \end{cases} \quad (2)$$

Тогда исходную плотность распределения $\varphi(x)$ можно представить в виде

$$\varphi(x) = \varphi_0(x) \sum_{i=0}^{\infty} c_i H_i(x). \quad (3)$$

Для определения коэффициентов данного разложения c_i необходимо (3) умножить на $H_j(x)$, проинтегрировать полученное равенство в пределах от $-\infty$ до ∞ и учесть условие ортогональности (2):

$$c_i = \frac{1}{i!} \int_{-\infty}^{\infty} \varphi(x) H_i(x) dx.$$

В частности, для центральных моментов имеем $c_0 = 1$, $c_1 = c_2 = 0$, $c_3 = \frac{1}{3!}\mu_3$, $c_4 = \frac{1}{4!}(\mu_4 - 3)$, следовательно, разложение (3) принимает вид

$$\varphi(x) = \varphi_0(x)\left[1 + \frac{1}{3!}\mu_3 H_3(x) + \frac{1}{4!}(\mu_4 - 3) H_4(x) + \ldots\right]. \quad (4)$$

Это разложение называется рядом Грама – Шарлье типа А [4]. На практике при использовании конечного отрезка ряда (4) для оценивания плотности по выборке конечного объема $u_1, u_2, \ldots, u_N$ центральные моменты заменяются оценками

$$\hat{\mu}_k = \frac{1}{N} \sum_{i=1}^{N} (u_i - \bar{u})^k.$$

Поскольку в качестве эталонной функции $\varphi_0(x)$ выступает плотность стандартного нормального распределения, то ряд (4) записан для центрированных и нормированных случайных величин. В общем случае для произвольной случайной величины X, имеющей математическое ожидание m и дисперсию σ^2, моменты μ_k в (4) необходимо заменить величиной μ_k / σ^k и использовать следующее разложение:

$$\varphi(x) = \frac{1}{\sigma}\varphi_0\left(\frac{x-m}{\sigma}\right)\left[1 + \frac{1}{3!}\frac{\mu_3}{\sigma^3} H_3\left(\frac{x-m}{\sigma}\right) + \frac{1}{4!}\left(\frac{\mu_4}{\sigma^4} - 3\right) H_4\left(\frac{x-m}{\sigma}\right) + \ldots\right]. \quad (5)$$

Учитывая, что $\beta_1 = \frac{\mu_3}{\sigma^3}$ – коэффициент асимметрии, а $\beta_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3$ – коэффициент эксцесса, получим

$$\varphi(x) = \frac{1}{\sigma}\varphi_0\left(\frac{x-m}{\sigma}\right)\left[1 + \frac{1}{3!}\beta_1 H_3\left(\frac{x-m}{\sigma}\right) + \frac{1}{4!}\beta_2 H_4\left(\frac{x-m}{\sigma}\right) + \ldots\right]. \qquad (6)$$

Как отмечается в [5], если $\varphi(x)$ имеет ограниченную вариацию, то ряд (6) сходится к $\varphi(x)$ в каждой точке непрерывности, а значит, может использоваться для аналитического представления $\varphi(x)$ с любой степенью точности.

Однако, как утверждается в [2], сходимость или расходимость ряда (6) не имеет практического значения, поскольку обычно ограничиваются небольшим количеством слагаемых (два, три). Поэтому важно, чтобы эти слагаемые позволяли с заданной точностью аппроксимировать искомую функцию плотности. Видимо, точность аппроксимации будет приемлемой в том случае, если эталонная функция $\varphi_0(x)$ выбрана достаточно удачно в смысле, что интересующее нас распределение $\varphi(x)$ близко к эталонному. Эта гипотеза требует более детальной проверки. Практика показывает [2], что большую часть встречающихся в реальности распределений удается представить отрезком разложения (6).

Алгоритм оценивания. Перейдем к построению алгоритма оценивания параметров уравнения регрессии (1) в предположении, что функция плотности распределения случайных ошибок представима в виде (6). В силу же предположения о независимости случайных ошибок уравнения (1) значения отклика y_i также будут статистически независимыми случайными величинами с плотностью распределения $\varphi(y_i)$, только $E[y_i] = x_i\theta$ (x_i – i-я строка матрицы X из (1)). Для оценивания параметров уравнения (1) можно воспользоваться методом максимального правдоподобия [2, 4]. Логарифмическая функция правдоподобия имеет вид

$$l(y_1, \ldots, y_N) = \ln\left(\prod_{i=1}^{N} \varphi(y_i)\right) = \sum_{i=1}^{N} \ln(\varphi(y_i)).$$

Учитывая разложение функции плотности (6) и ограничиваясь только приведенными слагаемыми, можно записать

$$l(y_1, \ldots, y_N) = l(y_1, \ldots, y_N, \theta) = \sum_{i=1}^{N} -\ln(\sigma\sqrt{2\pi}) - \frac{(y_i - x_i\theta)^2}{2\sigma^2} +$$

$$+ \ln\left(1 + \frac{1}{6}\beta_1\left(\frac{(y_i - x_i\theta)^3}{\sigma^3} - 3\frac{(y_i - x_i\theta)}{\sigma}\right) + \frac{1}{24}\beta_2\left(\frac{(y_i - x_i\theta)^4}{\sigma^4} - 6\frac{(y_i - x_i\theta)^2}{\sigma^2} + 3\right)\right).$$

Следует отметить, что максимизация данной функции проводилась численно с использованием метода вращающихся координат [6]. При этом необходимо постоянно обеспечивать положительность выражения, стоящего под знаком логарифма, для чего применялись барьерные штрафные функции [6].

Оценивание неизвестных параметров уравнения регрессии осуществлялось итерационно в соответствии со следующим алгоритмом.

Шаг 1. Определение некоторого начального приближения оценок вектора неизвестных параметров $\hat{\theta}^0$ ($k=0$), вычисленного, например, с помощью обычного метода наименьших квадратов (МНК) [1].

Шаг 2. Вычисление остатков уравнения регрессии и их центральных моментов до четвертого порядка включительно, а также коэффициентов асимметрии и эксцесса β_1, β_2.

Шаг 3. Поиск очередного приближения оценок вектора неизвестных параметров $\hat{\theta}^{k+1}$:

$$\hat{\theta}^{k+1} = \arg\max_{\theta} l(y_1, y_2, ..., y_N, \hat{\theta}^k).$$

Шаг 4. Если $\left\| \hat{\theta}^{k+1} - \hat{\theta}^k \right\| < \varepsilon$ (ε – заданная погрешность вычисления), то завершение процесса, в противном случае $k := k+1$ и переход на шаг 2.

Перед рассмотрением результатов вычислительных экспериментов, направленных на исследование предложенного метода, заметим, что случаев расходимости представленного алгоритма не наблюдалось. Видимо, это свидетельствует о достаточной степени аппроксимации плотности $\varphi(x)$ рядом Грама – Шарлье (6).

Результаты вычислительных экспериментов. Для исследования разработанного алгоритма оценивания вектора неизвестных параметров θ уравнения (1) проводились многочисленные вычислительные эксперименты. Приведем лишь некоторые из полученных результатов. В качестве истинной зависимости рассмотрим уравнение регрессии

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \varepsilon, \tag{7}$$

где количество регрессоров $p=3$, значения входных факторов x_{ij} выбирались из отрезка $[-1,1]$, истинные значения неизвестных параметров: $\theta_0 = 50$, $\theta_1 = 25$, $\theta_2 = 10$. Случайные ошибки ε_i моделировались независимыми и одинаково распределенными с функцией распределения

$$F(x) = (1-\lambda)F_1(x, m_1, \sigma_1) + \lambda F_2(x, m_2, \sigma_2), \tag{8}$$

где $F_i(x, m_i, \sigma_i)$ – функция нормального распределения с математическим ожиданием m_i и дисперсией σ_i^2, $i=1,2$; $\lambda \in [0,1]$ – параметр смеси. Во всех экспериментах $m_1 = m_2 = 0$ (если по тексту не оговорено противное).

Такое представление позволяет моделировать ошибку с различной степенью отклонения от нормального распределения, в том числе появление отдельных довольно грубых засоряющих наблюдений – «выбросов». Параметр λ определяет соответствующие доли наблюдений с дисперсиями σ_1^2 и σ_2^2 в выборке. При $\lambda = 0$ и $\lambda = 1$ ошибка будет иметь нормальное распределение. В проведенных вычислительных экспериментах полагалось, что $\sigma_2^2 > \sigma_1^2$. Однако при моделировании задавались не сами значения дисперсий

σ_1^2 и σ_2^2, а соответствующие им значения уровня шума. Уровень шума введен в [7] и определяется как отношение шум/сигнал (в процентах):

$$\rho = \frac{\sigma}{c} \cdot 100\,\%,$$

где σ^2 – дисперсия ошибки, $c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i^0 - \bar{y}^0)^2$ – интенсивность сигнала (незашумленных измерений y_i^0).

В качестве показателей точности оценивания параметров использовались L_1-нормы отклонений оценок неизвестных параметров от истинных значений

$$\psi_1 = \left\| \frac{\theta_{\text{ист}} - \hat{\theta}}{\theta_{\text{ист}}} \right\|, \quad \psi_2 = \left\| \theta_{\text{ист}} - \hat{\theta} \right\|.$$

Для различных комбинаций λ и ρ проводилось по 500 вычислительных экспериментов. Каждый такой эксперимент заключался в моделировании выборки исходных данных в соответствии с моделью (7) с последующим оцениванием параметров этой модели разработанным алгоритмом, а также МНК [1] и знаковым методом [8]. В качестве итоговых показателей точности оценивания использовались усредненные по 500 экспериментам значения показателей ψ_1 и ψ_2.

В табл. 1 представлены данные, полученные для нормально распределенной ошибки наблюдения ($\lambda = 0$) при разных объемах выборки. Дисперсия соответствовала уровню шума 5 %.

Из анализа результатов следует, что на малых выборках наилучшие результаты показывает МНК. Однако с увеличением объема выборки его преимущество перед разработанным алгоритмом становится незначительным. Это объясняется тем, что при увеличении объема выборки разложение Грама – Шарлье, а также асимптотическая эффективность метода максимального правдоподобия позволяют более точно аппроксимировать реальное распределение.

Таблица 1

Точность оценивания параметров уравнения регрессии при разных объемах выборки и нормально распределенной ошибке

Метод оценивания	Объем выборки N					
	50	100	200	50	100	200
	ψ_1			ψ_2		
МНК	0,0771	0,0575	0,0416	1,6996	1,2181	0,9059
Шарлье	0,0785	0,0589	0,0422	1,7507	1,2478	0,9181
Знаковый	0,0801	0,0606	0,0441	1,7985	1,2837	0,9517

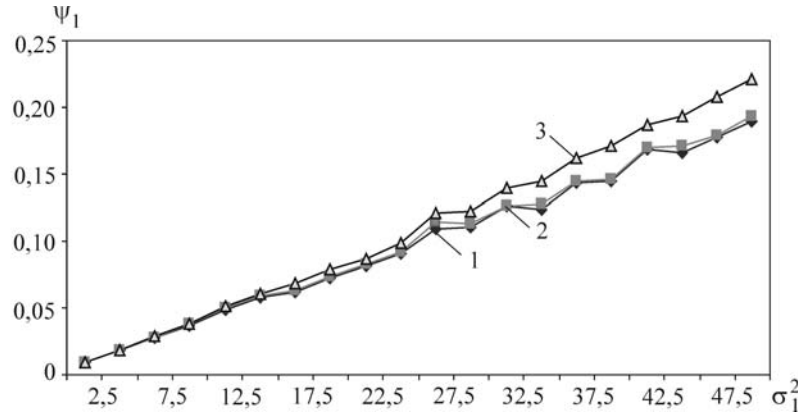

Рис. 1. Точность оценивания в зависимости от уровня шума при $N = 50$ (кривая 1 – МНК, 2 – метод Шарлье, 3 – знаковый метод)

Далее рассмотрим изменение точности оценивания в зависимости от дисперсии (уровня шума) нормально распределенной ошибки ($\lambda = 0$). Для этого будем последовательно изменять уровень шума, соответствующий дисперсии σ_1^2, от 2,5 до 50 % с шагом 2,5 (рис. 1). Из рисунка видно, что при малом уровне шума все три метода показывают примерно одинаковую точность. С увеличением уровня шума точность оценивания МНК и метода, основанного на аппроксимации (6), по-прежнему практически одинакова, в то время как точность оценивания знаковым методом стала несколько хуже, что отмечалось в [9].

Также проведено исследование точности оценивания вектора неизвестных параметров θ при разной степени отклонения распределения случайной ошибки от нормального распределения. Для этого изменению подвергался параметр смеси λ. При малых значениях λ в выборке будет появляться небольшое число выбросов, а при значениях λ, близких к 0,5, можно говорить

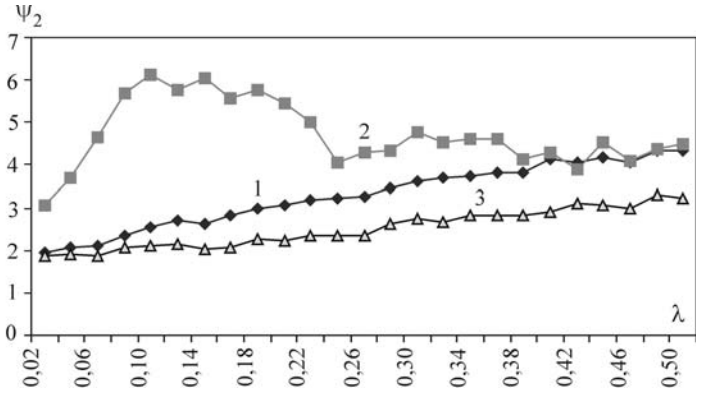

Рис. 2. Точность оценивания в зависимости от λ при $N = 50$ (кривая 1 – МНК, 2 – метод Шарлье, 3 – знаковый метод)

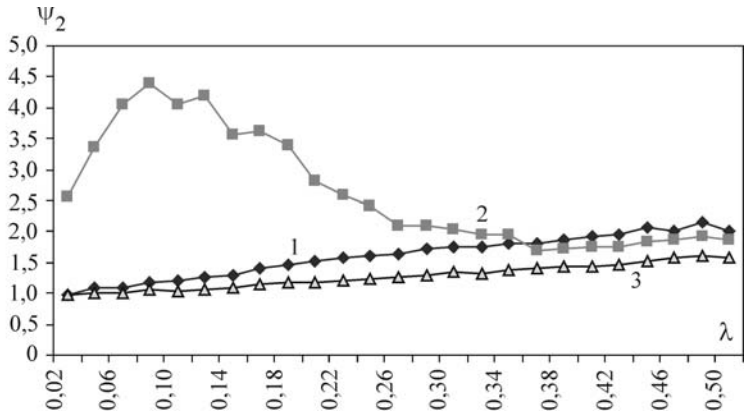

Рис. 3. Точность оценивания в зависимости от λ при $N = 200$ (кривая 1 – МНК, 2 – метод Шарлье, 3 – знаковый метод)

об изменении формы распределения. Было зафиксировано $\rho_1 = 5\ \%$, $\rho_2 = 50\ \%$, а доля выбросов λ изменялась от 0 до 0,5 с шагом 0,02. Результаты вычислительных экспериментов представлены на рис. 2 и 3.

Из рисунков видно, что оценивание параметров на основе разложения (6) дает хорошие результаты, если распределение ошибки существенно отлично от нормального (при больших значениях λ). Наличие небольшого числа выбросов, как правило, не меняет форму распределения, поэтому попытка оценивать распределение с помощью разложения (6) при малых λ не приводит к желаемому результату. В этом случае лучше использовать устойчивые к наличию выбросов методы, например знаковый метод [8] или метод наименьших уравновешенных квадратов (LTS – Least Trimmed Square) [10]. Преимущества знакового метода здесь можно объяснить сохранением при моделировании симметрии распределения ошибок.

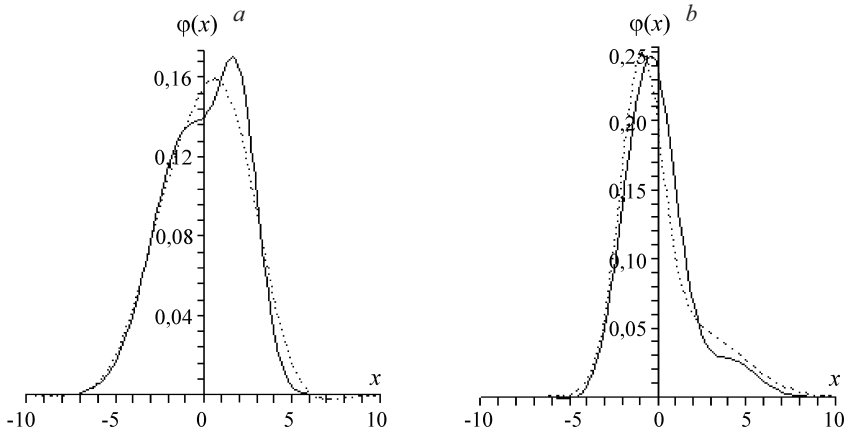

Рис. 4. Графики функции плотности (8): *a* – $\sigma_1 = 1{,}1$, $\sigma_2 = 2{,}0$; *b* – $\sigma_1 = 2{,}7$, $\sigma_2 = 1{,}2$

Таблица 2

Точность оценивания при ошибке, имеющей асимметрию и эксцесс, отличный от нормального распределения (N = 100)

Метод оценивания	ψ_1		ψ_2	
	$\sigma_1 = 1{,}1, \sigma_2 = 2{,}0$	$\sigma_1 = 2{,}7, \sigma_2 = 1{,}2$	$\sigma_1 = 1{,}1, \sigma_2 = 2{,}0$	$\sigma_1 = 2{,}7, \sigma_2 = 1{,}2$
МНК	0,0266	0,0269	0,5811	0,5881
Шарлье	0,0248	0,0229	0,5481	0,5211
Знаковый	0,0286	0,0258	0,6256	0,5912

Однако как только распределение случайных ошибок уравнения (1) имеет отличные от нормального распределения асимметрию или эксцесс, результаты становятся качественно иными. Например, положим в (8) $m_1 = 2$, $m_2 = -1$, $\sigma_1 = 1{,}1$, $\sigma_2 = 2{,}0$, $\lambda = 1/3$, что обеспечивает математическому ожиданию моделируемой случайной величины равенство нулю, и выберем σ_1 и σ_2 такими, чтобы получить различные варианты распределений, отличных от нормального. Графики функций плотности таких распределений представлены на рис. 4, *a*, *b* сплошными линиями.

Результаты оценивания параметров уравнения (7) тремя методами приведены в табл. 2. Очевидно, что при таком выборе распределения случайных ошибок их медиана равна нулю, что не нарушает предпосылки знакового метода [8]. Из табл. 2 видно, что наилучшие результаты показывает метод, основанный на разложении (6). Знаковый метод уступает даже МНК, что, видимо, связано с потерей симметрии распределения.

На рис. 4 пунктирными линиями показано восстановление функции плотности с помощью разложения (6) при объеме выборки 100 элементов. Это свидетельствует о хорошем качестве оценивания плотности и подтверждает преимущества данного подхода при оценивании параметров регрессионных зависимостей.

Заключение. На основе аппроксимации функции плотности случайных ошибок рядом Грама – Шарлье типа А в представленной работе предложен новый метод оценивания параметров регрессионных зависимостей. Использование информации, содержащейся в моментах более высокого порядка, а именно в коэффициентах асимметрии и эксцесса, позволяет говорить о возможности адаптации алгоритма оценивания к форме распределения случайных ошибок. Традиционно применяемый на практике метод наименьших квадратов не имеет такой гибкости, поскольку основан только на моментах первого и второго порядка (математическом ожидании и дисперсии). Проведенные вычислительные эксперименты показали, что в случае отклонения распределения случайной ошибки от нормального (появление асимметрии, изменение эксцесса) предложенный метод явно превосходит классический метод наименьших квадратов и знаковый метод. Последний рекомендуется к использованию при засорении выборки единичными выбросами.

СПИСОК ЛИТЕРАТУРЫ

1. **Рао С. Р.** Линейные статистические методы и их применения. М.: Наука, 1968.

2. **Пугачев В. С.** Теория вероятностей и математическая статистика. М.: Наука, 1979.
3. **Корн Г., Корн Т.** Справочник по математике для научных работников и инженеров. М.: Наука, 1984.
4. **Кендалл М., Стьюарт А.** Теория распределений. М.: Наука, 1966.
5. **Крамер Г.** Математические методы статистики. М.: Мир, 1975.
6. **Химмельблау Д.** Прикладное нелинейное программирование. М.: Мир, 1975.
7. **Ивахненко А. Г., Степашко В. С.** Помехоустойчивость моделирования. Киев: Наук. думка, 1985.
8. **Болдин М. В., Симонова Г. И., Тюрин Ю. Н.** Знаковый статистический анализ линейных моделей. М.: Наука, 1997.
9. **Денисов В. И., Тимофеев В. С.** Знаковый метод: преимущества, проблемы, алгоритмы // Науч. вестн. НГТУ. Новосибирск: Изд-во НГТУ. 2001. № 1(10). С. 21.
10. **Rousseeuw P. J., Van Driessen K.** Computing LTS regression for large data sets // Data Mining and Knowledge Discovery. 2006. N 12. P. 29.

Поступила в редакцию 31 января 2008 г.